KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering

# Characterisation and modelling of residential electricity demand

**Wouter Labeeuw**

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Engineering

December 2013

# Characterisation and modelling of residential electricity demand

**Wouter LABEEUW**

Supervisory Committee:
Prof. dr. ir. C. Vandecasteele, chair
Prof. dr. ir. G. Deconinck, supervisor
Prof. dr. T. Holvoet
Prof. dr. ir. R. Belmans

Prof. dr. ir. C. Develder
(Ghent University)
Prof. dr. rer. nat. R. Stamminger
(Universität Bonn)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Engineering

December 2013

D/2013/7515/155
ISBN 978-94-6018-774-2

# Preface

On March 13, 2009, I had a meeting with prof. G. Deconinck about the possibility of starting a Ph.D. at Electa. The topic, distributed control, had my immediate attention. The work itself was associated with a project called Linear, funded by IWT. While working on the Ph.D., the topic shifted away from distributed control towards data modelling, as it became clear that data and models were lacking to execute simulations.

Finding my way in research was challenging, but also fun and exciting. I would like to take this opportunity to thank some people who helped me throughout the Ph.D.

My supervisor Geert Deconinck for giving me the freedom to explore my ideas and for giving useful feedback. I always thought my work was still insufficient, until he told me: 'It is time to start writing the dissertation'. I very much appreciated his listening skills and his subtle ways of making me change direction by asking questions.

The members of the jury, prof. C. Vandecasteele for chairing and profs. T. Holvoet, R. Belmans, C. Develder and R. Stamminger for giving insightful comments and remarks on the first version of the dissertation. The work of prof. Stamminger also helped me in modelling appliances.

The people I worked with, colleagues at Electa and members of the Linear project, for their collaboration, the constructive meetings and for sharing their insights and ideas. The colleagues in the offices I worked, whichever, for the nice atmosphere, the entertaining coffee breaks and for sharing lunches. The secretary for helping us out with practical problems and for listening to our small aspirations. The technicians for their preparations and assistance during lab sessions. I couldn't have done it without you.

My friends and family, for ensuring me a good work-life balance. My parents, Hubert and Katelyn, for letting me do the things I wanted to do, for supporting me therein and for believing in me. And last but not least, my wife, Sofie, for being there during less pleasant times, for always believing in me and the love I received.

Thank you!

Wouter Labeeuw
Leuven, December 2013

# Abstract

Demand side management is regaining attention because of the integration of intermittent renewable energy sources such as wind and solar power. Accurate electrical load data are needed to estimate the potential and the impact of demand response. Data itself is hard to grasp. Analytical techniques are required to describe the data in a comprehensive way. Privacy laws prevent data usage in certain contexts. Data modelling gets around the privacy problem and allows for Monte Carlo simulations.

First, residential electricity demand is described. The demographic properties describing the electricity demand in Belgium are determined with the use of machine learning techniques. Residential electricity consumers are grouped based on timing of and amount of electricity demand with clustering techniques. Measurement data from a project are scaled up by employing the groups found to get an impression of electricity demand of wet appliances.

Then, residential electricity demand is modelled. The combination of electricity consumer groups together with Markov models allow for an electricity demand model. Wet appliance load cycles are detected and parametrised according to the previously defined consumer groups. The first is able to regenerate load data at the household connection point, the latter is able to regenerate load cycles of appliances.

Finally, the models are applied to estimate potential and impact of direct appliance control and corresponding privacy issues.

# Beknopte samenvatting

Het beheer van de vraagzijde komt meer onder de aandacht door de integratie van intermitterende hernieuwbare energiebronnen, zoals wind- en zonne-energie. Accurate data met betrekking tot elektrische last zijn nodig om het potentieel en de impact van vraagrespons in te schatten. De data zelf zijn moeilijk te interpreteren, daarom zijn analytische methoden nodig om data op een verstaanbare manier te beschrijven. Wetten met betrekking tot privacy beperken het gebruik van dergelijke data in bepaalde contexten. Het modelleren van de data vermijdt privacyproblemen en laat toe om 'Monte Carlo'-simulaties uit te voeren.

Eerst wordt de residentiële elektriciteitsvraag beschreven. De demografische eigenschappen die de residentiële elektriciteitsvraag in België bepalen, worden bepaald aan de hand van technieken voor machinaal leren. Residentiële elektriciteitsgebruikers worden met clusteringstechnieken gegroepeerd aan de hand van het tijdstip en de grootte van de elektriciteitsvraag. Metingen van een project worden opgeschaald, door gebruik te maken van de gevonden groepen, om een indruk te krijgen van de elektriciteitsvraag van wasmachines, droogkasten en vaatwassers.

Daarna wordt de residentiële elektriciteitsvraag gemodelleerd. De combinatie van de groepen elektriciteitsgebruikers, samen met markovmodellen, laten toe om een model van de elektriciteitsvraag op te stellen. De lastcycli van wasmachines, droogkasten en vaatwassers worden gedetecteerd en gegroepeerd en geparametriseerd aan de hand van de groepen elektriciteitsgebruikers. Het eerste model is in staat om lastprofielen van huishoudens te genereren. Het tweede kan lastprofielen van toestellen produceren.

Ten slotte worden de modellen toegepast om het potentieel en de impact van directe laststuring van bovenstaande toestellen in te schatten en de mogelijke privacyproblemen aan te duiden.

# Abbreviations

| | |
|---|---|
| ANN | Artificial Neural Network |
| API | Application Programming Interface |
| CV(RMSD) | Coefficient of Variation of the Root Mean Squared Deviation |
| DC | Direct Current |
| DER | Distributed Energy Resource |
| DSO | Distribution System Operator |
| DT | Decision Tree |
| EM | Expectation Maximisation |
| FKM | Fuzzy $k$-means |
| FN | False Negative |
| FP | False Positive |
| HTTPS | Hypertext Transfer Protocol Secure |
| ICT | Information and Communication Technology |
| IP | Internet Protocol |
| KDD | Knowledge Discovery and Data mining |
| KM | $k$-means, a distance based clustering algorithm |
| MLP | Multilayer Perceptron |
| NIALM | Non-intrusive appliance load monitoring |
| OLAM | Online Analytical Mining |

| | |
|---|---|
| OLAP | Online Analytical Processing |
| PNG | Portable Network Graphics |
| PPV | Positive Predicted Value |
| REST | Representational State Transfer |
| RMSD | Root Mean Squared Deviation |
| ROC | Receiver Operating Characteristic |
| SMO | Sequential Minimal Optimization |
| SOM | Self organizing maps |
| SQL | Structured Query Language |
| SSH | Secure Shell |
| SSL | Secure Sockets Layer |
| SVM | Support Vector Machine |
| SVN | Subversion |
| TLS | Transport Layer Security |
| TN | True Negative |
| TNO | Dutch Organization for Applied Scientific Research |
| TP | True Positive |
| TRIAC | Triode for Alternating Current |
| TSO | Transmission System Operator |
| VHF | Very High Frequency |
| VREG | Flemish Regulator for Electricity and Gas |
| WebDAV | Web Distributed Authoring and Versioning |

# List of Symbols

| | |
|---|---|
| $\mathbf{a}$ | Min-max normalised vector |
| $\alpha$ | Slope |
| $app$ | Appliance |
| $c$ | Specific heat coefficient |
| $cdf$ | Cumulative probability density function |
| $\chi^2$ | Pearson $\chi^2$ error |
| $cl$ | Class or Cluster |
| $cl(.)$ | Part of Class or Cluster, +1 if part of, -1 if not |
| $\mathbf{d}$ | Distance vector |
| $d$ | Dimension |
| $E$ | Electrical energy |
| $Err$ | Squared error |
| $\eta_{eff}$ | Efficiency |
| $Ex$ | Expected value |
| $f$ | Function |
| $g(.)$ | Output function of neural network |
| $G(.,.)$ | Information gain of a test |
| $i$ | Instance |
| $k$ | Weibull shape parameter |

| | |
|---|---|
| $K(.,.)$ | Kernel function |
| $K_l(.,.)$ | Linear kernel function |
| $K_p(.,.)$ | Polynomial kernel function |
| $K_r(.,.)$ | Radial basis kernel function |
| $L$ | Latent heat |
| $\lambda$ | Weibull scale parameter |
| $m$ | Mass |
| $\mathbf{m}_c$ | Cluster centre |
| $\mu$ | Mean |
| $N$ | Size of sampling frame |
| $n$ | Number of selected individuals |
| $O$ | Observed value |
| $\mathbf{P}$ | Transition matrix |
| $\mathbf{p}$ | Probability vector |
| $P$ | Power |
| $P(.)$ | Probability |
| $pdf$ | Probability density function |
| $\mathbf{q}$ | Vector representing start states of a Markov chain |
| $Q$ | Heat |
| $\mathbf{s}$ | Support vector |
| $s(S)$ | Entropy of set $S$ |
| $S$ | Set |
| $\mathcal{S}$ | Partition of sets |
| $\sigma$ | Standard deviation |
| $st$ | State of Markov chain |
| $T$ | Temperature |

| | |
|---|---|
| $t$ | Time |
| $\mathbf{v}$ | Vector representing weighted sum of inputs |
| $\mathbf{w}$ | Weight vector |
| $\mathbf{x}$ | Input vector |
| $X(t)$ | Stochastic process at time step t |
| $\mathbf{y}$ | Output vector |

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background

Smart grids are defined as reliable and more efficient electricity networks which allow for environment friendly electricity generation and distribution [1]. The implementation of smart grids is done by monitoring-, communication- and control technologies.

Linear [2] is a smart grid project that studies the possibilities to better align electricity demand and generation. The alignment is called demand response. Information related to changes in the market price and the electricity demand of the customer is presented to the customer for the demand response. Active demand goes one step further, the appliances are directly controlled, i.e. automatically started, switched off or their electricity demand is curtailed or expanded, within the boundaries set by the customer. The main difference with other European projects, such as ADDRESS[1] and EU-Deep[2] lays in the automation of demand response, the execution of a field trial, the inclusion of smart meters and a collaboration between a manifold of partners from industry and academia.

The research questions of the project are related to the technical and economical potential of active demand response. Simulations are executed to determine the possibility to shift, curtail or expand electricity demand of appliances and electrical vehicles and to calculate the grid impact thereof, requiring

[1] ADDRESS website: http://www.addressfp7.org
[2] EU-Deep website: http://www.eudeep.com

measurement data and models. The target of the project is the residential sector, a sector where privacy regulations are strict.

To technically or economically control appliances, knowledge of the residential electricity demand is required. Describing residential electricity demand helps to understand results of simulations. A selection of measurement data with related demographic information has to be made for specific simulations. Grouping customers based on electricity demand aggregates their data into representatives for the group, limiting the specific information of the individual customers. However, aggregated data is not always sufficient for simulations, therefore electricity demand is modelled, in a privacy friendly way, according to the found customer groups. The description, selection and modelling of residential electricity demand by using a data analysis approach is the focus of this thesis.

## 1.2 Data, information and knowledge

Data is situated at the bottom of the "data, information and knowledge"-hierarchy [3] and represents facts or observations. Information is interpretable and meaningful data.

The purpose of this thesis is to convert data into information and to add interpretation to the information when possible. Information describes the main properties of data and makes it easier for others to work with the data. The information systems help to interpret and may generate data. The interpretation is needed to understand the original data and to be able to draw conclusions.

Data is the starting point for simulations and decision making. However, data isn't always available because of privacy concerns or because of the possible commercial or strategic information they might provide. Data providing parties often opt to deliver aggregated instead of raw data. Aggregated data is useful in some studies, but when detail is important, aggregation can only be used up to a certain level or cannot be used.

## 1.3 Scope and objectives

To compare smart grid strategies and to find their impact, simulation need to be executed, requiring representative data to operate. Privacy laws prevent the use of residential data in certain contexts. The goal of the thesis is to describe, select and model residential electricity measurements and to check the potential for active demand response of the wet appliances in a privacy friendly way.

The focus is on residential electricity demand, more specifically the electricity demand at the connection point (the point of common coupling) of a house and the electricity demand of wet appliances.

Electricity demand in Belgium needs to be described based on demographic properties and based on groups of similar electricity consumption.

- Describe electricity demand and demographic parameters related to the total electricity demand of households in Belgium, making it possible to select customers in a non-representative set and to interpret results of simulations.
- Group households with a similar electricity demand pattern, in terms of timing and magnitude of demand, to create group representatives. Representatives consist of aggregated data, removing customer specific information. The grouping algorithm has to enable scaling up data, spreading customer information over the found groups.
- Determine wet appliances' electricity demand of those groups to get insights about how they use their appliances. Appliance electricity demand and usage is required to find the impact of appliances on residential electricity demand and to model appliances for simulations.

Measurements of the electrical power demand reveal information on households. Find privacy issues related to monitoring total electricity demand and the electricity demand of wet appliances. The questions to be answered are: 'Is it possible to find appliances' cycles in smart meter data (15' resolution)?' and 'What information can be derived from fifteen minute resolution appliance sub-metering data?'.

Residential load data is needed for simulations to estimate the technical and economical potential of active demand response. Information systems that select customers and model residential electricity demand based on the findings of the electricity demand description are needed.

- Customers interested in their electricity consumption over-respond in surveys and field test. Select a subset of customers from a biased set (due to non-response) based on both electricity demand and related demographic properties.
- Model electricity demand of households, enabling the generation of electricity demand of fictive households. Aggregating simulated load profiles of a customer group needs to result in the load of the group representative.

- Model electricity demand of wet appliances, enabling the generation of load cycles. Aggregating the simulated load cycles needs to again result in the wet appliances' electricity demand of the group representative.

Estimations about the potential for flexibility, the amount of electricity demand that can be shifted by the households that are willing to allow those shifts, based on the description of wet appliances' electricity demand are needed to determine the potential impact of active demand. The effect of using flexibility has to be assessed by using the models of the wet appliances, giving insights about what to expect from active demand response simulations.

An information system needs to be created to facilitate the above analyses. Its criteria are security and functionality, i.e. the ease of use.

## 1.4 Outline

Chapter 2 describes the data and the data sources for the analyses. The data consists of residential load measurements and survey data. The data sources are the Linear project and the distribution system operators Eandis[3] and Infrax[4].

Chapter 3 presents techniques and algorithms for selecting, describing and modelling. The 'knowledge discovery and data mining' process is explained. Machine learning techniques to fit and to describe data are elaborated upon. A short overview of Markov chains is given and sampling techniques are explained.

Chapter 4 gives an overview of the literature related to this thesis. The literature describing the following chapters is elaborated upon.

Chapter 5 describes the residential electricity demand in Belgium. A supervised machine learning algorithm is used to find the demographic parameters related to total electricity demand of households. An unsupervised machine learning algorithm groups households with a similar electricity demand. The wet appliance electricity demand is determined thereafter for those groups.

Chapter 6 discusses privacy issues related to measuring the electricity demand of households and their appliances. The most important factors are the resolution of the measurements. The working principle of wet appliances is explained and used to detect load cycles.

Chapter 7 explains how demand description is incorporated in models. Sampling load profiles from a dataset is done according to demographic parameters.

[3] http://www.eandis.be/
[4] http://www.infrax.be/

Electricity demand of households and the electrical load of wet appliances are modelled based on the groups of households with similar demand.

Chapter 8 presents the potential of shifting wet appliances and the impact shifting has on their demand. The electricity demand of wet appliances of the groups of households is combined with the attitude towards shifting to find the potential. The impact is tested with the help of models.

Chapter 9 deals with software infrastructure to perform the analyses described above. The infrastructure handles security, while the software adds the functionality.

# Chapter 2

# Data description

An overview of data and data sources required for the analyses is presented. First, the difference between load curves and load profiles is explained. The synthetic load profiles of the regulator, used to describe the residential electricity demand, are elaborated upon. A set of residential load profiles representative for Flanders are the basis of the synthetic load profiles. Such a load profile represents the electricity demand over a year of a household. The measured households are interviewed for a survey to determine their demographic properties. Finally, the data of the Linear project is described. It contains measurements to calculate first estimates for active demand, as well as preliminary measurements of the field test.

## 2.1 Load curves vs. profiles

A load profile is the variation of electrical power per time step for a period in time. The higher the resolution, i.e. the smaller the time step, the larger the load profile becomes. A common time step for measuring electrical power by distribution system operators is fifteen minutes, also being the time resolution used in the balancing market. A typical load profile describes the active power demand of a year in a detailed way. Other durations are possible as well. The load profile of a customer is the electricity demand measured at the connection point of the house. For the remainder of the dissertation, load profiles consist only of load, excluding embedded generation. Load profiles can also be measured for appliances and districts.

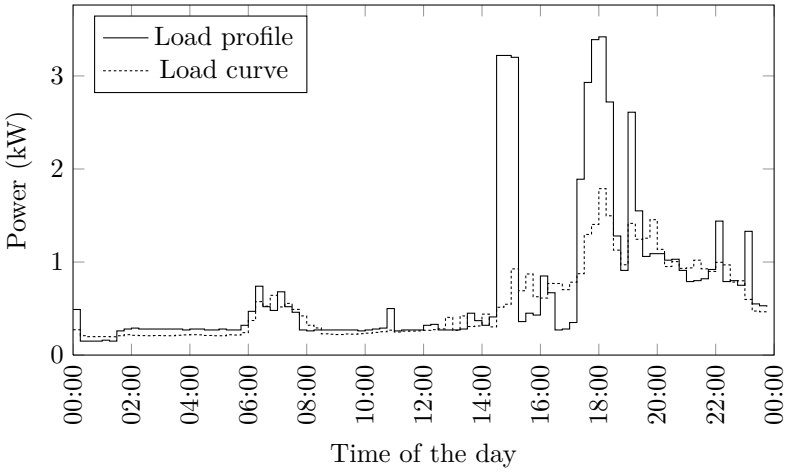



Figure 2.1: Load profile and load curve

Load curves are aggregated forms of load profiles. A load curve depicts the average power demand for a given period in time. The aggregation is done over different days. An example of a load curve is the electricity demand of an average day in January.

Figure 2.1 shows the difference between a load profile and a load curve. The load profile is the power demand of a customer at a Monday in February of 2008. The load curve represents the power demand at the average Monday in February of 2008. The load curve is a smoothed version of the load profile: peaks are lower, but the trend is very similar. A load curve captures the general behaviour of the customer, while a load profile represents the detailed behaviour. The disadvantage of load curves is the inability to represent load of customers with large variations on power demand an the timing thereof.

Load curves are a way of handling data gaps. If a load profile misses a power measurement, the gap is not taken into account to determine the load curve. The same result would be obtained if the missing value is replaced by the average value of the other power measurements at that time of the day. Replacing missing values with averages is a common way to handle them [4].

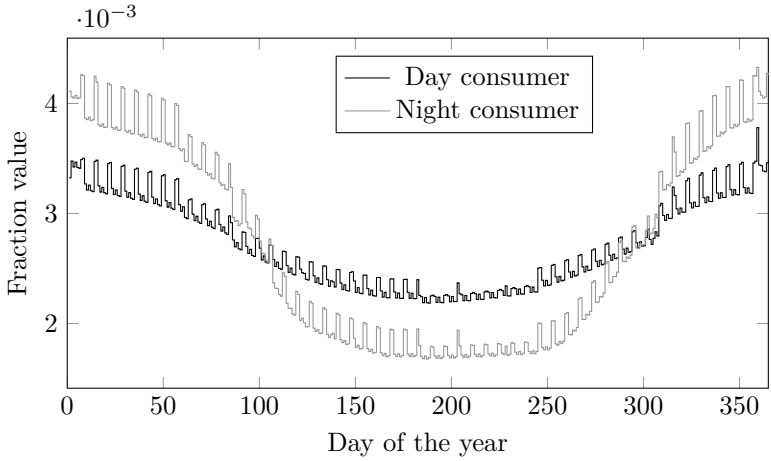



Figure 2.2: Synthetic load profiles of the residential electricity demand for 2012

## 2.2 Synthetic load profiles

Synthetic load profiles (SLP) describe the general electricity demand of groups of customers. The profiles are used in the settlement process between various market parties when no measurement data of the customer is available. An SLP consists of a series of 35040 numbers or fractions. The fractions describe how the total electrical energy demand at a connection point is spread over a year. The resolution is fifteen minutes and the sum of all the fractions is one [5].

The electricity demand during the year of a customer is estimated by multiplying the total yearly electricity consumption of the customer with the fractions. The confidence level of a synthetic load profile is 95 % with a precision of 5 % [5]. A synthetic load profile is not representative for one, but for a group of customers. The impact on a distribution grid of one household or the potential for demand response cannot be determined with a synthetic load profile, but requires measured load profiles.

Six types of synthetic load profiles are defined for the electricity market: two residential, two non-residential, a flat consumer and public lighting [6]. The focus is on the residential sector, only those types are discussed.

The synthetic load profiles for electricity are based on measurements at the connection point of a representative set of at least 2400 industrial and residential customers [5]. Measurement data of over 1200 residential customers (1363 in 2008) are the basis for the residential electrical synthetic load profiles.

The two residential types are night and day consumers. Night consumers have an electricity demand during the night and weekends that is higher than or equal to 1.3 times the demand during the day ($E_{night}/E_{day} \geq 1.3$). The other customers are day consumers ($E_{night}/E_{day} < 1.3$). Figure 2.2 shows both synthetic load profiles. Seasonal effects are clearly visible. The electricity demand in January is more than double (2.29 times) that in July for the night consumer. The difference between winter and summer is less explicit for the day consumers, where electricity demand in January is 1.48 times the demand during the July.

The distinction of day and night consumers is because of the tariff schemes in Belgium. The tariffs schemes are day tariff (between 6h and 21h or 7h and 22h, depending on the distribution system operator), night tariff (during the weekend and when not day) and exclusive night tariff (8 to 9 hours during the night, only for electric heating and hot water purposes). A flat tariff also exists with a price in between day and night tariff.

A season is a quarter of a year long. Databases are not designed to work with seasons. However, databases have functionality to work with quarters of a year. The first quarter year starts at the first of January and ends at the last day of March. The period differs 11 days from winter, 22nd of December until 21 of March. Considering the small difference in days and the ease to work with quarters of a year, quarters are preferred over seasons in this thesis.

## 2.3 Residential load data of Flanders

Synthetic load profiles of electricity are calculated on a set of over 1200 residential customers, as explained in Section 2.2. The same set of data forms the basis for analysis. The data are provided by the distribution system operators Eandis and Infrax via the Linear project [7, 2].

The dataset consists of a total of 1693 measured connection points for electricity, spread around Flanders, the Dutch speaking northern part of Belgium. 1363 of the 1693 connection points are measured in 2008. Every year, some households are dropped from the set and others are added. At each connection point, the average fifteen minute power is measured [8].

The installed base of small (< 10 kW) photovoltaic (PV) systems was only around 15080 at the end of 2008 [9]. With 2.6 million households in Flanders: about 0.6 % of the total population had a PV installation, hence load profiles are regarded as electricity demand only. The number of small PV installations increased a lot afterwards, to 216516 (8.3 %) in July 2012 [10].

Table 2.1: Age distribution of total population, sample and weighted sample

| Age | Population [%] | Sample [%] | Weighted sample [%] | Sample size |
|---|---|---|---|---|
| 18 - 34 | 23.4 | 4.8 | 14.0 | 20 |
| 35 - 44 | 19.1 | 10.3 | 15.3 | 43 |
| 45 - 55 | 19.0 | 23.0 | 21.1 | 96 |
| 56 - 64 | 15.6 | 28.2 | 20.5 | 118 |
| $\geq 65$ | 22.9 | 33.7 | 29.1 | 141 |

## 2.4 Survey data

Surveys are a useful method to describe the characteristics of a population. The requirements are a probability sample and a data collection method. The list of households measured for the synthetic load profiles, is the sampling frame (i.e. the set to sample from). 1326 households of Flanders are called upon via letter and telephone to make an appointment for a face-to-face interview. 500 households responded [11], 416 of which have been measured in 2008.

The data collection method is a questionnaire conducted in-home. The questions are close-ended: the respondent only has to indicate the appropriate answers. The questionnaire includes demographic, building, mobility, insulation, energy demand, heating, ICT, appliance, ecological and directly controlled appliance related questions [11].

An over-representation of certain population groups was found in the survey data. The demographic properties 'number of inhabitants' and 'age' are correlated to electricity demand (Section 5.1). In the survey, elderly persons and households of two persons over-responded (Tables 2.1 and 2.2) [12]. To adjust for the over-representation, the sample is weighted according to both 'age' and 'number of inhabitants'. The combined weight is the average of both. Weighing each member relative to the joint probability of age and number of inhabitants would be more suitable, however the Belgian Directorate General for Statistics and Economic Information does not provide such information.

Each sample member is weighted according to its combined 'age' and 'number of inhabitants' weight. In comparison to the original sample, the weighted sample is closer to the real population. The description of demographic parameters of the survey in the rest of the text refers to the weighted sample.

Table 2.2: Distribution of the number of inhabitants of the total population, the sample and the weighted sample

| Inhabitants | Population [%] | Sample [%] | Weighted sample [%] | Sample size |
|---|---|---|---|---|
| 1 | 29.8 | 17.5 | 22.3 | 73 |
| 2 | 34.2 | 47.9 | 36.6 | 200 |
| 3 | 15.8 | 17.9 | 19.4 | 75 |
| 4 | 13.7 | 10.8 | 14.2 | 45 |
| $\geq 5$ | 6.6 | 6.0 | 7.6 | 25 |

## 2.5 Project data

The source for appliance data is the 'Linear'-project. The electricity demand at the connections point and at appliance level is measured for a set of households. Amongst the different metered appliances are washing machines, tumble dryers, dishwashers, freezers, fridges, boilers, ovens and air-conditioning systems. The focus for appliances is on washing machines, tumble dryers and dishwashers because those appliances are considered to be good candidates for active demand [13]. They have a relatively high maximum power, are easily controllable (start or delay) and have a high penetration level.

Two types of project data are used: measurements for the first estimates and field-test data. The measurement data for the first estimates contains appliances and measurements at the connection point of the home. The field-test data has the same types of measurements and contains action loggings such as appliance configuration. Additional households are monitored for the field-test. Three months of field-test data is available when writing the thesis text.

### 2.5.1 Dataset 1

The measurement data for the first estimates is called 'dataset 1' and consists of measurements at the connection point of households and of appliances. Measurements of both are only available for a limited set, in most cases only appliances are measured.

56 households are measured at the connection point, of which only 42 are considered reliable. The reliability criteria are: at least half a year of measurement points and the measured power should be at least 8 W (i.e. the power consumption of the measurement unit). A wireless connection (WiFi) is used for communication between measurement unit at connection point and home gateway [14]. The home gateway is a device that collects measurement data

Table 2.3: Number of reliable measurements series for appliances

| | total | load curve | load curve & conn. point |
|---|---|---|---|
| Washing machine | 88 | 63 | 30 |
| Tumble dryer | 69 | 50 | 27 |
| Dishwasher | 53 | 41 | 21 |

sends the data to the back-end system. The measurement units at connection point do not have a buffer. Each time the connection fails, measurements aren't tracked. An extra reliability criterion is the possibility to make a load curve of the average week for each quarter of the year. In this way, seasonal effects are captured. A load curve of the average week for each quarter of the year could be constructed for 39 of those 42 households.

The appliance measurements are much more reliable [15]. The connection between the measurement unit at the appliance and the home gateway is provided by Zigbee [14]. A buffer at the measurement unit ensured no data got lost. However, some households unplugged the measurement units for some period in time. 88 washing machines, 69 tumble dryers and 53 dishwashers were equipped with a metering unit. For 63 washing machines, 50 tumble dryers and 41 dishwashers, a load curve of the average week for each quarter of the year could be constructed.

Reliable measurements at connection point and at appliance level combined, allow for a hierarchical view on electricity demand. The combination of measurements at connection point and appliance level are available for 30 households with a washing machine, 27 households with a tumble dryer and 21 households with a dishwasher. An overview of the total number of measured appliances, the number of appliances with enough data for load curves and number of appliances with both enough data for load curves at appliance level and at connection point is presented in Table 2.3.

### 2.5.2 Dataset 2

The field-test data contains data from March 3, until June 20, 2013 and is called dataset 2. The electricity demand at the connection point and of various appliances, together with the appliance's configurations are logged. The measurements at the connection point aren't used because of the limited duration of the measurements: no load curves of the average days of the week for the quarters of the year could be made. The appliances' measurement data

together with the appliances' configuration logging enable to validate appliance cycles and appliance settings detection algorithms.

The appliance logging data consists of the selected programme, a time stamp of when the programme was selected and a deadline for the appliance's start. The latter is required for the active demand of the field test. The programme settings couldn't be logged because of compatibility issues.

The field test has an installed base of 169 washing machines, 154 tumble dryers and 120 dishwashers, which are logged. The measurement data is checked for cycles of the appropriate appliances by verifying whether the appliances are used and whether the profile matches the appliance. 104 dishwashers, 98 tumble dryers and 67 dishwashers are detected in the data up to June 20, 2013.

## 2.6 Conclusions

An overview of the data sources is given. The differences between load curves and load profiles are explained. The residential electricity demand is presented from the point of view of the regulator. The data used by the regulator is described and will be used further on. A large part of the households that participated in the measurements of the regulator, responded to a survey. A description of the survey and the answers to the survey is given. Also project data is used in the thesis. The measurements for the first estimates and the field-test measurements are described.

# Chapter 3

# Knowledge discovery and representation

Knowledge discovery and representation refer to selection, analysis and interpretation of data. The 'knowledge discovery and data mining'-process describes how to get useful knowledge out of a data set. The process is the basis for machine learning. Both supervised as unsupervised machine learning techniques are explained. Markov chains are a special way to represent a random process: describing transitions between states. The sampling section gives more information on ways to select data. The methods are introduced as they have been considered or used in the subsequent chapters.

## 3.1 KDD process

The concept of 'Knowledge discovery in databases' dates from the so-called workshop of the 'International Joint Conference on Artificial Intelligence' in 1989 [16]. However, because of the incorporation of data mining in the process and the naming of journals such as 'Advances in Knowledge Discovery and Data Mining', the KDD acronym is also referred to as 'Knowledge Discovery and Data mining'. Nevertheless, the KDD process is still defined as [17],

> The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

The term 'KDD' is mainly used to refer to the process of useful knowledge discovery from data sets [18]. The KDD process depicted in Figure 3.1 has the different steps [17]

**data selection,** creating a target set, a subset of the data,

**data preprocessing,** removing noise and outliers, handling missing and unlabelled data,

**data transformation,** finding features more efficient to represent the data, usually involving dimension reduction and data projection,

**data mining,** choosing and applying an appropriate algorithm to find patterns and relationships in data,

**interpretation,** consolidating discovered knowledge.

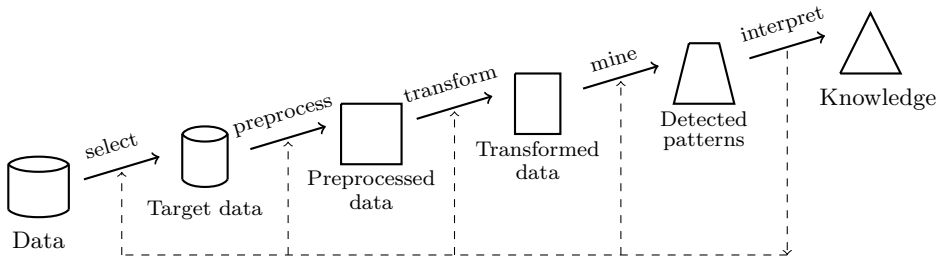



Figure 3.1: Knowledge discovery and data mining (KDD) process

The KDD process allows for iterations between the different steps: if the outcome of a certain step is insufficient, changes in previous steps can be made.

The data mining step relies on machine learning and/or statistics. Statistical analysis assumes a stochastic model for the data to predict or describe the data. Machine learning on the other hand uses an algorithm that operates on the data to predict or describe the data [19].

## 3.2 Machine learning

Machine learning is the capacity of a computer program to improve its performance at solving a problem after receiving information about the problem [20], or more formal [21]

> Given a task $T$, a performance criterion $C$, and experience $E$, a system learns from $E$ if it becomes better at solving task $T$, as measured by criterion $C$, by exploiting information in $E$.

The learned model is a function $f : \mathbf{x} \rightarrow \mathbf{y}$ which predicts the outcome $\mathbf{y}$ based on input $\mathbf{x}$. Machine learning algorithms are subdivided into supervised, unsupervised and semi-supervised learning. Supervised learning algorithms learn function $f$ based upon given $\mathbf{x}$ and $\mathbf{y}$ vectors. Unsupervised learning algorithms on the other hand learn only from $\mathbf{x}$, the $\mathbf{y}$ vectors are not given. Semi-supervised learning is in between supervised and unsupervised learning: $\mathbf{y}$ is given for certain instances, but not all of them.

### 3.2.1 Supervised learning algorithms

The focus of supervised learning algorithms is prediction rather than description. The learned function or model is considered a generalisation of mapping between input and output. When a new input is presented, the function is able to predict the outcome [20].

Within supervised learning, three algorithm groups stand out: artificial neural networks (ANN), support vector machines (SVM) and decision trees (DT). The most common algorithms ([4]) of those groups are selected and compared. Multilayer perceptron (MLP) is the algorithm of choice for artificial neural networks. For support vector machines, sequential minimal optimization (SMO) is chosen. C4.5 [22] is selected as decision tree algorithm.

#### Artificial neural networks

The principle of artificial neural networks is based on biological networks of neurons. A biological neuron has different inputs. If the nonlinear 'sum' of the inputs is higher than a threshold, the neuron 'fires'. An artificial neuron is the mathematical description of a biological neuron: it takes the weighted sum of its inputs and applies the neuron activation function $g$ to the result. The result of the activation function is 1 (fire) or $-1$. An artificial neural network is composed of a number of interconnected neurons.

Figure 3.2 shows how neurons are interconnected in case of a multilayer perceptron. The connections between layers of a multilayer perceptron are directed, which makes the structure a feedforward artificial neural network [20, 4]. Layer 1 is the input layer, layer 2 is the hidden layer and layer 3 is the output layer in Figure 3.2. $y$ is the final result. Each connection between neurons is

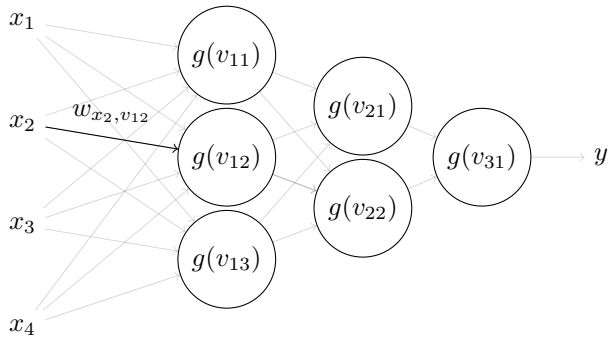



Figure 3.2: Multilayer perceptron

weighted. The weighted sum is expressed as $v_{ln}$, with $l$ as layer number and $n$ the number of the neuron in the layer. Equation 3.1 gives $v_{12}$ as example.

$$v_{12} = \sum_{i=1}^{n_x} w_{x_i,v_{12}} \cdot x_i \tag{3.1}$$

A step function is required to get a 1 or $-1$ as output of a neuron. However, gradient descent is used to determine (learn) the weights of the neurons, which makes it impossible to use a step function because of its indifferentiability. A sigmoid function (3.2) is an approximation of a step function, with a simple differential (3.3). The substitution of $v$ by the weights $w$ and inputs $x$ results in the differentiation to $w$ (3.4).

$$g(v) = \frac{1}{1+e^{-v}} \tag{3.2}$$

$$\frac{dg(v)}{dv} = \frac{0-e^{-v}}{(1+e^{-v})^2} = \frac{1}{1+e^{-v}} \cdot \frac{(1-e^{-v}-1)}{1+e^{-v}}$$

$$= g(v) \cdot (1-g(v)) \tag{3.3}$$

$$\frac{dg(v)}{dw_i} = \frac{dg(v)}{dv} \cdot x_i \tag{3.4}$$

The direction in which a function propagates is called the gradient of the function and is equal to the vector of partial derivatives of the function. The

partial derivative of the squared error ($E$) according to the weights is

$$Err = \frac{1}{2} \cdot (y - g(v))^2 \tag{3.5}$$

$$\frac{dErr}{dw_i} = (y - g(v)) \frac{dg(v)}{dw_i} \tag{3.6}$$

Substitution of Equations (3.3) and (3.4) in (3.6) shows how the weight is adapted according to changes in the error

$$\frac{dErr}{dw_i} = (y - g(v)) \cdot g(v) \cdot (1 - g(v)) \cdot x_i \tag{3.7}$$

The change in error is not directly applied to the weight, a step size $s$ controls the speed of change in the weights

$$\Delta w_i = s \cdot \frac{dErr}{dw_i} \tag{3.8}$$

The above reasoning assumes just one neuron. Updating weights for multiple layers of neurons is similar. The equations of the output layer are substituted into the equations of the hidden layer and those equations are substituted into the equations of the input layer. The algorithm to change the weights is called backpropagation.

Training of multilayer perceptrons goes as follows. The weights of the neurons are first initialized randomly. For each instance, i.e. learning example, the backpropagation algorithm is executed. If the output error is still too large, the algorithm takes the first instance again until the error is reduced sufficiently or until the maximum number of iterations is reached.

### Support vector machines

Perceptrons, as presented above, try to find a hyperplane which separates positive (1 or 'fire') and negative ($-1$) instances. More than one hyperplane might satisfy the separation of training instances. However, the hyperplane might not perfectly separate unseen instances. Support vector machine

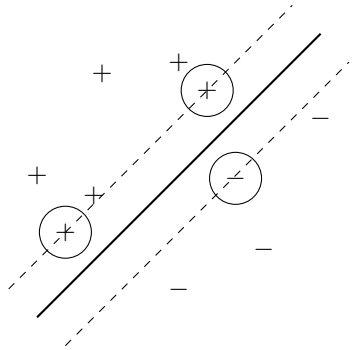


Figure 3.3: Training examples, support vectors (circled), linear hyperplane (line) and margins (dashed line)

algorithms try to maximise the distance between hyperplane and training instances. The closest positive and negative training instances are kept at the same distance from the hyperplane and are called support vectors. Figure 3.3 illustrates training instances, support vectors, hyperplane and equidistant margins.

The simplest form of hyperplane is a linear one. The output $y$ is expressed as a linear combination of the support vectors $\mathbf{s}$ and the input $\mathbf{x}$

$$y = b + \sum_{i=1}^{n_{\mathbf{s_i}}} \alpha_i cl(\mathbf{s_i}) \mathbf{s_i} \cdot \mathbf{x} \tag{3.9}$$

Each support vector $\mathbf{s_i}$ has a corresponding output $cl(\mathbf{s_i})$ which is 1 or $-1$. $b$ determines the offset of the hyperplane and $\alpha_i$ the slope corresponding to the support vector. The support vector machine algorithm tries to maximise the margin between hyperplane and support vectors, by adjusting offset and slope parameters. The linear combination of support vectors and input vectors can be described as a linear kernel $K_l$.

$$K_l(\mathbf{s}, \mathbf{x}) = \mathbf{s} \cdot \mathbf{x} \tag{3.10}$$

thus

$$y = b + \sum_{i=1}^{n_{\mathbf{s_i}}} \alpha_i cl(\mathbf{s_i}) K_l(\mathbf{s_i}, \mathbf{x}) \tag{3.11}$$

Support vector machines use this linear model to implement non-linear hyperplanes by replacing the linear kernel function by a polynomial

$$K_p(\mathbf{s}, \mathbf{x}) = (\mathbf{s} \cdot \mathbf{x})^n \tag{3.12}$$

In general, the degree $n$ of the polynomial function is not known in advance. The support vector machine algorithm starts with a linear kernel and increases the degree until the distance between training instances and the hyperplane ceases to increase. Another possible kernel is a radial basis function

$$K_r(\mathbf{s}, \mathbf{x}) = e^{-\frac{||\mathbf{s}-\mathbf{x}||^2}{2\sigma^2}} \tag{3.13}$$

The support vector machine algorithms adjusts the standard deviation $\sigma$ of the Gaussian distribution in order to determine the correct hyperplane.

Sequential minimal optimization is an iterative algorithm for optimizing the kernel parameters to determine the hyperplane separating positive and negative instances.

### Decision trees

A decision tree is a graph mapping inputs $\mathbf{x}$ onto an output $y$. Decision trees consist of internal nodes and leaves. A leave is an end node and contains an output value. Multiple leaves allow for multiple output values. Internal nodes have a test function (yes/no questions), the outcome of the test determines the next node to go to (Figure 3.4).

Decision trees are built using a divide-and-conquer approach. A root node, i.e. the top node of the tree, is selected first. The training instances are divided amongst two branches of the root node. New nodes are placed at the end of the branches, again subdividing training instances. Once a pure subset of the data is found, the procedure stops for that branch and a leaf node is placed at the end.

The measurement of purity in decision trees is entropy. Entropy is the number of bits needed to represent missing information. The entropy $s(S)$ of a set $S$ is

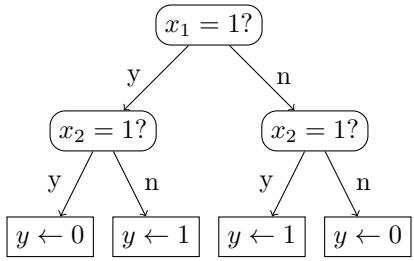



Figure 3.4: Decision tree representing an xor function

calculated based on the probability $p$ of each class $cl$.

$$s(S) = -\sum_{c=1}^{n_{cl}} p(cl) \log_2 p(cl) \tag{3.14}$$

A partition $\mathcal{S}$ of sets $S_s$ has an average class entropy which is the weighted average of the entropies of its sets.

$$s(\mathcal{S}) = -\sum_{s=1}^{n_s} \frac{|S_s|}{|S|} \sum_{c=1}^{n_{cl}} p_s(cl) \log_2 p_s(cl) \tag{3.15}$$

The information gain $G$ of a test $\tau$ is the difference between the entropy of the total set $S$ and the average entropy of the partition $\mathcal{S}$ of sets $S_s$ created by the test $\tau$.

$$G(S, \tau) = s(S) - s(\mathcal{S}) \tag{3.16}$$

A decision tree algorithm calculates the information gain of multiple tests $\tau$. The test with the largest information gain is selected and becomes an internal node. The sets $S_s$ resulting from the test are placed at the branches of the newly created node.

The above described decision tree algorithm is called ID3. C4.5, the algorithm used to determine the properties that predict demand, is based on ID3, but adds methods to work with numerical input, missing values and noisy data [4]. C4.5 is considered to be a benchmark algorithm [20].

### 3.2.2 Unsupervised learning algorithms

The main goal of unsupervised learning is to describe a data set. Clustering is the most prominent unsupervised learning algorithm and refers to the task of finding patterns in a data set [20].

Clustering algorithms group various instances based on a similarity or alternatively a dissimilarity measure. Dissimilarity measures are easier to define as they can be expressed as a distance metric $d$, where the outcome is a positive real number ($d : \mathbf{x} \times \mathbf{x} \rightarrow \mathbb{R}^+$).

The first distinction within clustering algorithms is the difference between flat and hierarchical clustering. In flat clustering, clusters are formed on a single level and are disjoint. Hierarchical clustering on the other hand divides the data set in different levels, into clusters with sub-clusters.

$k$-means (KM) and Expectation Maximisation (EM) clustering are described in this section. The considered algorithms are flat clustering algorithms, because hierarchical clustering is prone to outliers [23], as explained in Section 4.1.1.

#### $k$-means clustering

$k$-means (KM) clustering is an algorithm which groups instances based on the Euclidean distance between them. The number $k$ is the only parameter of the algorithm and refers to the resulting number of clusters the algorithm has to find. Data normalisation before clustering ensures a better performance [24].

The clustering algorithm works iteratively. First (at step $t = 0$), $k$ instances (i.e. training examples) are randomly selected. The selected instances $cl$ are called seeds or cluster centres. An empty set $S$ is associated with each seed. For each new instance $i$ in the data set, the Euclidean distance $d$ between the point $\mathbf{x}_i$ and each seed $\mathbf{m}_{cl}$ is calculated,

$$d_{i,cl}^{(t)} = \left\| \mathbf{x_i} - \mathbf{m_{cl}^{(t)}} \right\| \tag{3.17}$$

Each instance $\mathbf{x}_i$ is added to the next $(t+1)$ set $S$ of the closest seed $cl$,

$$S_{cl}^{(t+1)} = \left\{ \mathbf{x_i} : d_{i,cl}^{(t)} \leq d_{i,j}^{(t)}, \forall 1 \leq j \leq k \right\} \tag{3.18}$$

The seeds are updated for the next step and receive the value of the mean of the set:

$$\mathbf{m_{cl}^{(t+1)}} = \frac{\sum\limits_{x_k \in S_{cl}^{(t+1)}} \mathbf{x_k}}{\left|S_{cl}^{(t+1)}\right|} \tag{3.19}$$

The process of calculating distances, adding to seed sets and updating the seeds of the set is repeated until all instances remain within their set. An optimal clustering is not guaranteed with KM clustering, the initial seeds influence the resulting clusters, but it works well in general [20].

### Expectation Maximisation clustering

The Expectation Maximisation algorithm is an iterative method for learning mixture models [20], representing the distribution of a population by a mixture of various sub-population distributions. EM clustering tries to estimate the likelihood that an instance belongs to a sub-population. Each sub-population is a cluster.

The Expectation Maximisation clustering algorithm requires knowledge of the number of clusters $n_c$. KM clustering starts at an initialisation step in the EM clustering algorithm. After initialisation, the algorithm alternates between its two main steps:

**Expectation,** calculate for each instance $i$ the probability of belonging to the different clusters,

**Maximisation,** calculate the cluster centres mean and standard deviation.

Bayes' theorem is the basis of the algorithm. The probability of an instance $i$ belonging to a cluster $cl$ is

$$P(S_{cl}|\mathbf{x}_i) = \frac{P(S_{cl}) \cdot P(\mathbf{x}_i|S_{cl})}{P(\mathbf{x}_i)} = P(\mathbf{x}_i \in S_{cl}) \tag{3.20}$$

where $P(S_{cl})$ is the probability of the cluster (the prior), $P(x_i|S_{cl})$ the likelihood of $x_i$ given $S_{cl}$ and $P(x_i)$ as a normalizing constant, defined by,

$$P(\mathbf{x}_i) = \sum_{j=1}^{n_c} P(S_j) \cdot P(\mathbf{x}_i|S_j) \tag{3.21}$$

The model to describe the data distribution is a Gaussian distribution, calculated for each dimension $d$ of the data set. The likelihood of an instance $i$ being taken from the Gaussian distribution of a cluster $cl$ in dimension $d$ is defined by the probability density function value $pdf$ which relies on the mean $\mu$ and the standard deviation $\sigma$ in the dimension $d$ of cluster $cl$,

$$pdf_{i,cl,d} = \frac{1}{\sqrt{2\pi{\sigma_{cl,d}}^2}} e^{-\frac{(x_{i,d}-\mu_{cl,d})^2}{2{\sigma_{cl,d}}^2}} \tag{3.22}$$

To make the calculation numerically more stable, the log-likelihood is calculated,

$$\log pdf_{i,cl,d} = -\log\sqrt{2\pi} - \log\sigma_{cl,d} - \frac{x_{i,d}-\mu_{cl,d}}{2{\sigma_{cl,d}}^2} \tag{3.23}$$

Assuming Naive Bayes, i.e. the individual dimension probabilities are independent, the overall likelihood of instance $i$ belonging to cluster $cl$ is obtained by multiplying the different probability density values,

$$P(\mathbf{x}_i|S_{cl}) = \prod_{d=1}^{n_d} pdf_{i,cl,d} = \exp\left(\sum_{d=1}^{n_d} \log pdf_{i,cl,d}\right) \tag{3.24}$$

The numerator of Bayes' theorem (3.20) is called the density and can be rewritten as,

$$P(S_{cl}) \cdot P(\mathbf{x}_i|S_{cl}) = \prod_{d=1}^{n_d} pdf_{i,cl,d} \cdot P(S_{cl}) = \exp\left(\sum_{d=1}^{n_d} \log pdf_{i,cl,d} + \log P(S_{cl})\right) \tag{3.25}$$

The logarithm of the density is,

$$\log dens_{i,cl} = \sum_{d=1}^{n_d} \log pdf_{i,cl,d} + \log P(S_{cl}) \tag{3.26}$$

The EM algorithm can be executed as classifier (hard) or fuzzy (soft). In the hard case, the instance is assigned to the cluster with the highest likelihood, i.e. the probability of the most likely cluster equals one. The (soft) probability of belonging to a cluster is calculated by combining (3.25), (3.21) and (3.20). To make the calculation numerical more stable, the probabilities are normalized.

Equation (3.25) is replaced by subtracting the maximum log density from all densities,

$$P(S_{cl}) \cdot P(\mathbf{x}_i|S_{cl}) = \exp\left(\log dens_{i,cl} - \arg\max_{cl} \log dens_{i,cl}\right) \tag{3.27}$$

The Gaussian models are updated during the Maximisation step. Each cluster $cl$ has a Gaussian model in every dimension $d$, which consists of a mean $\mu$,

$$\mu_{cl,d} = \frac{\sum_{i=1}^{n_i} P(\mathbf{x}_i \in S_{cl}) \cdot x_{i,d}}{\sum_{i=1}^{n_i} P(\mathbf{x}_i \in S_{cl})} \tag{3.28}$$

and a standard deviation $\sigma$,

$$\sigma^2_{cl,d} = \frac{\sum_{i=1}^{n_i} P(\mathbf{x}_i \in S_{cl}) \cdot (x_{i,d} - \mu_{cl,d})^2}{\sum_{i=1}^{n_i} P(\mathbf{x}_i \in S_{cl})} \tag{3.29}$$

The algorithm stops iterating between Expectation and Maximisation when the overall likelihood stops changing significantly between iterations. Significance is expressed as a threshold:

$$\left(\sum_{k=1}^{n_k}\sum_{cl=1}^{n_{cl}} \log dens_{k,cl}^{(t+1)} - \sum_{k=1}^{n_k}\sum_{cl=1}^{n_{cl}} \log dens_{k,cl}^{(t)}\right) \leq threshold \tag{3.30}$$

The threshold is defined experimentally at $10^{-10}$ for 10 successive iterations.

## 3.3 Markov chains

A discrete-time stochastic process describes the evolution of random variables $\{X(0), X(1), \ldots, X(t), \ldots\}$, where $X(t)$ is the value of the system characteristic at time $t$. [25] A Markov chain is a type of discrete-time stochastic process, where the outcomes of the random variables are a set of finite states [25].

Markov chains consist of states and transitions between them. Each transition has a certain probability. Figure 3.5 depicts a Markov chain with two states, $A$ and $B$. The probability of the transition from $A$ to $B$ is written as $P(B|A)$ is read as the probability of $B$ given $A$.

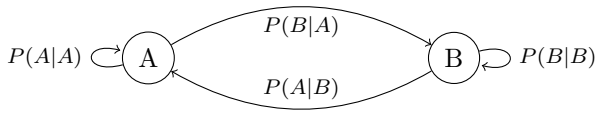



Figure 3.5: Two state homogeneous first order Markov chain

A Markov chain fulfils the Markov property: the next state $X(t+1)$ is sampled from a distribution $P(X(t+1)|X(t))$ depending only on the current state $X(t)$, [25, 26]

$$P(X(t+1) = i_{t+1}|X(t) = i_t, \ldots, X(1) = i_1, X(0) = i_0) \quad = \quad P(X(t+1) = i_{t+1}|X(t) = i_t) \tag{3.31}$$

The distribution $P(X(t+1)|X(t))$ is called the transition kernel. A Markov chain is time-homogeneous if the distribution $P(X(t+1)|X(t))$ remains the same at each time step. [26] The probability to go from state $i$ at time step $t$ to state $j$ at time step $t+1$ for homogeneous Markov chains is called the transition probability and has value $p_{ij}$,

$$P(X(t+1) = j|X(t) = i) = p_{ij} \tag{3.32}$$

The transition probabilities are described by an $n_{st} \times n_{st}$ transition matrix $P$, with $n_{st}$ the number of states,

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n_{st}} \\ p_{21} & p_{22} & \cdots & p_{2n_{st}} \\ \vdots & \vdots & & \vdots \\ p_{n_{st}1} & p_{n_{st}2} & \cdots & p_{n_{st}n_{st}} \end{bmatrix} \tag{3.33}$$

For every state $i$ at time $t$, there is a state at time $t+1$:

$$\sum_{j=1}^{n_{st}} P(X(t+1) = j|X(t) = i) = 1 \tag{3.34}$$

or

$$\sum_{j=1}^{n_{st}} p_{ij} = 1 \tag{3.35}$$

The probabilities of the start states $P(X(0) = i) = q_i$ are defined by the initial probability distribution of the Markov chain,

$$\mathbf{q} = \begin{bmatrix} q_1 & q_2 & \dots & q_{n_{st}} \end{bmatrix} \tag{3.36}$$

Sampling states from a Markov chain is done by randomly picking a transition based on its probability. The selected state is the starting position to select the next state.

The above description explains first-order, homogeneous Markov chains. In non-homogeneous Markov chains the transition probabilities change in time. Higher order Markov chains include memory. For example, the state of the next step in a second-order Markov chain depends on the current and the previous step $P(X(t+1)|X(t), X(t-1))$; $X(t-1)$ is kept in memory.

## 3.4 Sampling

Sampling is a technique to select individuals from a population. The purpose is to study the sample, obtain results and summary statistics, generalised for the whole population. Only part of a population is required to study the whole. The more individuals included in the sample, the better the summary statistics are [27, 28].

The population the sample is taken from, is called the sampling frame, with size indicated by $N$. The samples are the selected individuals, number $n$ [27, 28].

Two different approaches to sampling exist: probability and non-probability sampling. Probability sampling uses random selection and allows statistical theory for examining the results. Non-probability sampling is cheaper and more convenient, but does not allow for statistical theory [29, 27, 28].

### 3.4.1 Probability sampling

Probability sampling uses random selection. Each individual has a known non-zero probability of being included. These techniques are theoretically sound [27].

### Simple random sampling

Simple random sampling is a scheme with the property that each individual from the total population has the same probability of being selected. A simple way to sample randomly is to assign a random number (computer generated) to each individual, sort the population by the random number and pick the first $n$ individuals. The method cannot be used if the individuals are not labelled prior to sampling [27, 29, 28].

### Systematic sampling

To reduce the effort of selecting individuals, systematic sampling is applied. Each individual is numbered from 1 to $N$. The interval between draws $k$ equals $N/n$. Starting from individual $j < k$, each individual which equals $j + (a \cdot k)$ (with $a$, a number from 0 to $n - 1$) is selected [27, 28].

### Stratified sampling

Stratified sampling requires stratification of the population, i.e. the population needs to be divided into $L$ mutually exclusive and exhaustive strata. In general, a uniform sampling fraction is used: the sampling size of a stratum is proportional to the corresponding population stratum, called proportionate stratification. Disproportionate stratification is used to maximise the precision of the population mean estimator. The sampling fraction is then chosen to be proportional to the standard deviation of the stratum. The advantage of stratification is a greater precision (lower variance) compared to random sampling [27, 29].

### Cluster sampling

Cluster sampling requires all individuals to be part of one cluster, just as in stratified sampling. The difference between stratified and cluster sampling lies in the selection of individuals. Stratified sampling requires sampling from all strata, while with cluster sampling, the whole cluster is selected. The method is cheaper than other random sampling techniques, but has a higher sampling error [27, 29].

#### Multi-stage sampling

Multi-stage sampling uses combinations of the above sampling schemes. Multi-stage cluster sampling uses a hierarchy of clusters: first large clusters are selected, small ones within them and so on. An example of cluster sampling is selecting certain schools, within those schools select classrooms and select all students from the classroom. The multi-stage cluster sampling can be combined with simple random sampling where students are randomly sampled [27, 29].

### 3.4.2 Non-probability sampling

Non-probability sampling does not rely on random selection. The major disadvantage of non-probability sampling is that no statistical theory to examine the properties of the sampling estimators can be developed. Despite this theoretical weakness, non-probability sampling is widely used because of cost and convenience [29, 28].

#### Convenience sampling

Convenience sampling is also called accidental or haphazard sampling. Individuals are sampled by coincidence, examples being 'man on the street' interviews and patients available for a clinical trial. There is no evidence that they are representative of the population [29, 28].

#### Modal instance sampling

The mode of a distribution is the most frequently occurring value. Modal instance sampling uses this information to sample the 'typical' or the most frequently occurring case. The problems with modal instance sampling are the knowledge, or the not knowing, what the mode of the distribution is and the representativeness of the modal individual [28].

#### Heterogeneity sampling

Heterogeneity or diversity sampling is the opposite of modal instance sampling: diversity rather than typical individuals are selected. The purpose is getting a broad spectrum of individuals, i.e. including a large number of outliers. The selection of outliers is useful when extremes need to be tested [28].

### Snowball sampling

Snowball sampling is used to reach populations otherwise inaccessible or hard to find, not to find a representative sample. First, an individual meeting the selection criteria, is selected. The person is asked to recommend (select) other individuals who also meets the criteria, who are on their turn encouraged to find persons that meet the criteria [28].

### Expert sampling

A 'representative' sample is chosen by one or more experienced persons. In practice, different experts do not always agree on what constitutes a 'representative' sample. The bias of the survey estimators increases with the sample size: higher sample sizes lower the variance of a probability sample estimator, while the bias of the expert sampling estimator decreases only little [29, 28].

### Quota sampling

The two different types of quota sampling are proportional and non-proportional quota sampling. A minimum number of sampled individuals are required for non-proportional quota sampling: there have to be enough individuals of each quotum, the selection of more individuals in comparison to the quotum is not a problem. Proportional quota sampling on the other hand requires that the selected number of individuals is proportional to the quotum. The quota are a way to reduce the selection bias. However, the bias of the survey estimators increases with the sample size [28].

Quota groups are related to strata: both representing population groups from which samples are taken. However, quota sampling requires distributions over the whole population whereas stratification sampling requires mutual exclusive groups of the population [29].

## 3.5 Conclusions

An overview of the techniques required to perform the analyses in the thesis is given. First, the 'knowledge discovery and data mining'-process is explained. The process is the basis for machine learning. The supervised learning techniques 'support vector machines', 'artificial neural networks' and 'decision trees' are

elaborated. $k$-means clustering and Expectation Maximisation clustering are the explained unsupervised machine learning techniques. Markov chains are described and are used to model load profiles. The various sampling techniques are enlisted in the final part of the chapter.

# Chapter 4

# Literature review

An overview of the literature related is given in this chapter. The literature is used as basis, as a way to identify gaps in research and to position the work. Each section is related to a specific chapter in the thesis. The topics in the literature review handle electricity demand description, appliance detection and privacy, load profile selection and modelling, the potential for active demand management and tools to perform analyses on electricity data.

## 4.1 Demand description

The description of electricity demand helps to understand how people consume electricity. Understanding electricity demand enables to make statements on electricity consumption. Electricity demand at the connection point of a house as well as the electricity demand of wet appliances are described.

### 4.1.1 Electricity demand at connection point

Electricity demand at the connection point is described by two methods: demographic parameters influencing total annual electricity demand and grouping similar customers.

The literature review of demographic parameters mainly focuses on Dutch research, as it is considered closest to the Belgian/Flemish situation. The studies are based on questionnaires.

The algorithms to group customers focus solely on electricity demand. No specific country is targeted in the literature overview of electricity clustering.

### Demographic properties

Research [30] from 2001 concludes that appliance ownership and usage together with surface area and household size determines electricity demand. The ownership and the frequency of using a washing machine, a dryer, a dishwasher, electrical cooking plates and a pump for central heating are of most determining regarding appliances. Households size and surface area are less determining for electricity demand than appliance ownership.

A later study indicates that the residential electricity demand increased between 1990 and 1995 due to a lower number of persons per household (and hence a larger number of households), a larger hot water demand per person, more electric drying, freezing and cooling and a higher penetration of appliances [31]. Simultaneously, research is published which relates energy demand with net household income, household expenditure, age and the number of persons per household [32].

De Groot et al. describe an internal report of TNO in which number and age of occupants, duration of being at home, income, shower and bath usage frequency, heating and ventilation behaviour, use of devices and motivation to save energy are regarded important factors [33].

Mansouri et al. found that income, appliance ownership, persons per household and occupancy patterns were of importance for electricity consumption [34]. Another UK study suggests that floor area, number of occupants, number and type of appliances and occupancy patterns are key [35].

### Groups based on electricity demand

Electricity demand can also be described by grouping similar electricity demand patterns. The Flemish Regulator for Electricity and Gas (VREG) describes six groups of electricity consumers of which two are residential. The groups are defined using a statistical analysis [5].

Cluster analysis is, next to statistical analysis, a popular way to define customer groups. The most common algorithms to cluster electrical loads are $k$-means (KM) [36, 23, 37, 38], fuzzy $k$-means (FKM) [36, 23, 37, 38], hierarchical [36, 23, 37], modified-follow-the-leader [36, 23] and Expectation Maximisation

(EM) [39] clustering, self organizing maps (SOM) [36, 23, 38, 40, 41, 42] and ISODATA [43].

Chicco et al. [36, 23] compare KM, FKM, hierarchical, modified follow-the-leader and SOM clustering. Hierarchical and modified-follow-the-leader clustering score better in validity indicators (cluster dispersion indicator, modified Dunn indicator, scatter index and Davies–Bouldin index) than KM and FKM [36, 23]. Both are better at isolating outliers, resulting in smaller cluster dispersion and reduced scattering. KM and FKM keep the cluster population relatively uniform and detect outliers to a much lesser extent [23]. Zhang et al. [38] compare k-means, FKM and SOM and found that KM performed slightly better than FKM in the stability index. FKM on its turn performed better than SOM. Coke et al. [39] pointed out that mixture models are better in smoothing random effects and used a modified Expectation Maximisation clustering to group electrical load series.

Various data transformation techniques for load profiles are described in the literature. The most common transformations are harmonics analysis [44] and principal component analysis [45], Fourier series analysis [41], representative load patterns (i.e. normalised load profiles of one day) [46, 36, 23] and normalised load curves [38, 37, 42]. The normalisation techniques are dividing the powers of the load profile by the maximum value [46, 36, 23, 42], Z-score normalisation [38] and min-max normalisation [37].

Load curves are according to the number of publications the most common ones. A load profile represents detailed electricity demand information and randomness, as explained in Section 2.1. A load curve on the other hand describes the behaviour or pattern of electricity demand. Grouping customers with a similar behaviour or pattern is hence preferred.

### 4.1.2 Appliances

The electricity demand of wet appliances is described extensively by Stamminger et al. [47]. The sources for the work are European [48, 49] and German [50] studies. A representative load cycle profile (15 minute resolution) is created for various appliances. The load cycle profiles are coupled with the probabilities of starting during a certain period of the day and the average number of cycles during a week. The result is a load curve of the wet appliance electricity demand.

The start probabilities for an appliance are derived from a survey of 2500 customers from 10 countries. Customers could indicate the frequency of starting the appliances during different periods during the day. The periods during the day are ‘morning’, ‘midday’, ‘afternoon’, ‘evening’ and ‘night’. Frequencies vary

between 'always' (100%), 'often' (75%), 'sometimes' (50%), 'rarely' (25%) and 'never' (0%). The selected frequencies are thereafter normalised per household, and smoothed by calculating the moving average over three hours.

### Washing machine

An average electricity consumption of 0.89 kWh per load cycle is assumed for washing machines in [47]. The load cycle profile is based on measurements [47, 50]. The interpolation of two studies [50, 48] resulted in a total annual electricity demand of 150 kWh per year per household, i.e. approximately 170 cycles per year.

Washing machines are most often started in the morning, in the late afternoon and in the early evening. The average power demand is lowest during the night (14 W on average) and highest (approximately 52 W) in the morning and the early evening.

### Tumble dryer

The start probabilities of the tumble dryers are assumed to be the same as the ones for washing machines, but shifted 2 hours in time. A load cycle profile with an electricity demand of 2.46 kWh is assumed. The value is calculated by dividing the total annual electricity demand from tumble dryers in fifteen European countries [49] by the total number of households in the region, by assuming a penetration level of 34.4 % and that 60 % of the washing cycles are accompanied by a tumble dryer cycle.

The average power demand during the night is lowest (40 W). Two peaks of approximately 150 W are found. The peaks occur two hours after those of the washing machine.

### Dishwasher

The load cycle profile of dishwashers is determined in a similar way. The total electricity demand of dishwashers in fifteen European countries is divided by the total number of households of those countries, resulting in 251 kWh per household per year. The average number of cycles per week (4.06) is determined by a questionnaire [48]. Each load cycle profile requires 1.19 kWh per cycle.

During the night, the average power demand is lowest (27 W). The average power demand rises during the afternoon until the peak after dinner at 20h. The average power demand during the peak is 87 W.

## 4.2 Privacy

Cambridge Dictionaries defines privacy as 'someone's right to keep their personal matters and relationships secret'. When data is collected from people, it is harder for them to do so.

Distribution system operators are encouraged by governments to implement smart meters [51] to measure the electricity demand of households and communicate the information. Dutch law researchers already pointed out possible privacy issues [52]. Home energy management systems are getting popular as well, which means more data collection.

### 4.2.1 Appliance detection

Non-intrusive appliance load monitoring techniques allow detection of appliances in the measured electricity demand at the connection point of the household. With home energy management systems, no algorithms are required as appliances are measured individually.

The use of appliances and other detailed electricity demand provide insights in the lifestyle. Data is stored at the distribution system operator or at the provider of the home energy system. The customer hence loses control over their own data.

#### Non-intrusive Appliance Load Monitoring

The detection of appliances at the connection point of a house is called non-intrusive appliance load monitoring (NIALM), introduced by Hart [53], who used 1 Hz monitored data to detect appliances by active power. The method got improved by adding median filters [54].

Another method that works on a 1 Hz resolution is presented by Zeifman [55]. A probabilistic approach is used to detect on/off appliances such as toasters, washing machines, etc. Finite state appliances (e.g. a fan with multiple rotation speeds), variable power devices (e.g. dimmers) and permanent devices (always on) aren't detectable by this approach.

Lisovich et al. [56] are able to detect large distinctive appliances in a 1 or 15s resolution data. However, the overall ability to identify appliances is limited because of the high number of false positives and the low detection rate. 'Presence and sleep' schedules of the occupants are highly detectable.

In fifteen minute data, Powers et al. [57] are able to detect high load appliances such as water heaters and air-conditioners using appliance specific rules, considering the time of use and the power demand to determine the correct appliance.

Farinaccio et al. [58] work with a lower measurement resolution: 16s. Large devices such as refrigerators, washing machines and dishwashers are detected with appliance specific decision rules. The downsides of the approach is the need for one week training and appliance specific rules. The method got improved [59] using pre-processed changes in active power instead of signatures.

Another attempt to detect appliances in 15 minute basis data is made by Kolter et al. [60]. A sparse coding algorithm is used to train basis functions and activities of k classes of appliances (a class per type of appliance). The combination of the basis functions and the activities results in the electricity demand of the appliance class. To detect appliances, the reverse logic is applied: the activities of appliances are determined from the load at the connection point of the household using the basis functions. The electricity demand of an appliance class is reconstructed by combining the activities of the appliance class with the basis functions. However, the algorithm is only able to achieve at most 59 % classification accuracy.

Other ways to detect appliances use more than only active power or have a high measurement resolution. Harmonics and Fourier transformations are amongst the most popular ones [61, 62, 63]. I-V curves [64] and combinations of other techniques [65] are also used. A more detailed overview of NIALM methods can be found in [66]. The main conclusion of Zeifman [66] is that there isn't a complete NIALM solution for all types of appliances.

### Submetering

No non-intrusive appliance load monitoring is suitable to monitor all types of appliances as mentioned earlier. Therefore, submetering, i.e. metering every appliance, is considered to be a viable solution.

Ueno et al. [67] shows that energy demand information systems yield an average reduction in electricity demand by 9 %. Their home energy management system used appliance specific measurements. Park et al. [68] use smart sockets, socket watt meters with communication capabilities, to create a simulation framework

that models a network of home appliances and smart meters. Han et al. [69] propose a home energy management system with smart sockets using ZigBee and infra-red remote controls.

### 4.2.2 Concerns

Lisovich et al. [56] state that particularly law enforcement agencies, marketing firms and criminals are interested in in-home data. According to Quinn [70], there are seven concern types related to privacy: nefarious uses, insurance adjustments, target marketing, inquiries regarding disputes, inquiries regarding regulated activities, discrimination and profiling and medical questions. McKenna et al. [71] reduced the groups of privacy concerns to: illegal, commercial, law enforcement, other party legal purposes and co-inhabitants.

The illegal use mainly focusses on burglaries: checking whether a house is unoccupied [70, 56, 72] and detecting the possession of certain appliances [56]. Commercial usage is split in marketing and insurance companies and mainly considers the lifestyle of the inhabitants [72]. Advertisers are interested in milestones of a person's life, as they involve major changes in buying behaviour, amongst their milestones becoming a parent, moving homes, getting engaged and going through a divorce [73]. Eating habits, TV and computer use and whether or not you sleep well [70], together with the detection of specific appliance brand and malfunctioning appliances are also of interest for advertisers [56]. Insurance companies would like to know your lifestyle: do you arrive at home when bars close, do you get enough sleep, are appliances left on when leaving the house [70].

The detection of illegal activities and the verification of claims in court, e.g. was the person at home during the evening, are useful for law enforcement [56]. Other legal uses are an employer who checks if the employee is at home when reported to be sick [70]. Co-inhabitants checking each others behaviour [74] is also a use of smart metering data which compromises privacy,

## 4.3 Electricity demand for simulations

The description of electricity demand mostly focusses on aggregated data. However, for some problems, more detailed data is required. Section 4.3.1 describes techniques to select customers in pilot projects. Section 4.3.2 focusses on how electricity demand is modelled.

### 4.3.1 Selection

The preferred way to sample customers for a pilot project is using random sampling, because it allows for a comparison between control and treatment groups [75]. A sufficient number of customers is needed to draw conclusions. A trade-off between the information value and the cost of sampling has to be made.

A Swedish study selected 500 households as part of a time of use pilot [76]. The study didn't indicate how the households were selected. A set of 40 households was selected in a demand response pilot project in Norway [77]. The participants were recruited by means of a local news paper. The voluntary basis for participation suggests that people interested in demand response, over-responded.

Random selection is very costly (Section 3.4). The approach taken in [77] is convenience sampling. The technique is easy and less costly than simple random sampling, but comes with an over-representation of people willing to participate in such a project and who are interested in demand response. This problem is also addressed in the 'Linear' project (Section 7.1).

### 4.3.2 Modelling

Two strategies exist to model residential low voltage electricity demand: bottom up and top down. Bottom up approaches model electricity demand of each individual appliance. The sum over all appliances results in the electricity demand at the connection point. Top down approaches on the other hand start with the electricity demand at connection point and try to model it without knowledge of appliances.

The advantage of bottom up approaches compared to top down is that the electricity demand of the various appliances is known. Demand response and active demand strategies can be tested more easily. The downsides of bottom up approaches are the intensity of modelling and the risk of missing appliances to model [78].

#### Bottom up

Capasso et al. [79] base their household load model on two types of functions: 'behavioural' and 'engineering'. The 'behavioural' functions include a probability of being at home, home activities, appliance ownership, appliance use and human resources. Appliance's mode of operation, the maximum power at connection

point and technological penetration are 'engineering' functions. The sum of the individual loads of the appliances make up the electrical power demand. The time resolution is 15 minutes.

A dataset of 1200 homes where 175 households were equipped with appliance monitoring infrastructure, is the basis of the load modelling done by Stokes [80]. The electricity demand during the year of each appliance is modelled by sinus-functions. The appliance use is scaled by a weight determined by the number of inhabitants. Daily patterns ensure that appliances are started at the appropriate time. Appliance load cycles with a resolution of 1 minute are used to build electricity demand. The washing machine model for example uses 3 load cycles selected at random each time a washing machine is started. Both active and reactive power are modelled.

Energy demand (thermal and electrical) in households have two determinants according to Yao et al. [81]: behavioural and physical. The approach is hence similar to Capasso et al. [79]. The composition of the household and the occupancy pattern together with appliance ownership and usage constitute the behavioural determinant. The physical determinants are energy consumption of the various appliances and the energy consumption of domestic hot water and heating.

Widén et al. [82] use activity schemes to model electricity and hot water demand. An activity describes what a person is doing, vacuum cleaning for instance. The model for daily household activities is based on a diary of 431 persons in 169 households, filled in for one weekday and one weekend day between August and December by each person. An average power demand (on an hourly basis) is associated with each activity. The results are validated against two data sets: five households measured in a 10 minute interval during autumn and the aggregation of a dataset of 217 submetered households.

Markov chains are used to model the occupancy in households [83]. The occupancy models are combined with daily activity patterns to generate switch-on events [84]. The events trigger appliance load cycles. Most appliances have an assumed constant power, except for those with time-varying demand such as washing machines. No explicit explanation of the specific load cycles of time-varying appliances is given. The sum of the power demand of all appliances is the power demand at the connection point of the dwelling. The model is tested against 22 households, shows a slight underprediction of the variation between dwellings, an annual mean daily demand comparable to typical household profiles and an aggregated load duration curve similar to that of the measured profiles.

#### Top down

McLoughlin et al. have a top down approach to model residential electrical power demand [85]. The electricity load profiles of five households are modelled by a homogeneous Markov chain. The model is able to recreate the distribution of the electrical power. However, the autocorrelation, common in load profiles, could not be reproduced.

## 4.4 Flexibility

Flexibility in this thesis refers to the amount of shiftable or curtailable electrical load. The term is related to demand response, which is described by the U.S. Department of Energy as [86],

> Changes in electric use by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized.

However, lower electricity use is not essential for demand response. A more recent definition, from the European University Institute, focusses more on the active part and goes as follows [87],

> Changes in electric usage implemented directly or indirectly by end-use customers/prosumers from their current/normal consumption/injection patterns in response to certain signals.

The driving factors and uncertainties with respect to demand response are described first. An overview of studies and field tests is given thereafter. The attitude towards active demand has been investigated in the Linear project.

### 4.4.1 Driving factors and uncertainties

Various reasons exist for the implementation of demand response. The advantages are related to better operating markets, reduction of blackouts and a more efficient grid use. The downsides of demand response are mostly related to costs.

More efficient markets and competitive electricity markets, together with a better allocation of economic resources are the main advantages of demand response according to Kirschen [88]. The uncertainties are related to costs. New equipment is required to facilitate demand response. The installation of the equipment and the modification of the production schedules will often not be cost efficient according to Kirschen [88]. The recurring communication cost of metering systems for smaller consumers will often outweigh the potential economic savings.

Saele et al. [77] mention the avoidance of blackouts as one of the main reasons to implement demand response. However, consumers see electricity as a 'low interest' product and might not be interested in demand response. The cost for the installation of the demand response equipment is also a concern in [77].

The advantages of demand side management according to Strbac [89] are less back-up generation capacity, a higher efficiency of the transmission grid, less transmission and distribution grid investments and the balancing of renewables. The lack of ICT infrastructure connected to the power grid is considered to be a disadvantage. Demand side management is also not competitive to traditional approaches in most cases. The complexity of operating the grid will increase for the system operators.

The main concerns for policy makers are the limited knowledge of potential savings and expected costs [90]. Customers also lack means and incentives to participate in demand response [90]. A comprehensive representation of the price and the incentives is therefore required [91]. The incentives will have to be high enough for them to care [91, 92].

The lack of bidirectional communication, an ICT problem, was a problem in New Bern, North Carolina, USA. 30 % of their switches to directly control load failed over the years: the utility did not know which appliances were connected [93].

### 4.4.2 Studies and field tests

Early demand response programmes focussed on the industrial and the commercial sector [94]. Álvarez Bel et al. [95] divided the commercial sector into categories in order to identify the potential for demand response. Universities, hospitals, shopping malls, hotels and large offices were regarded as most promising.

The interest of customers in green electricity in Canada is tested by Rowlands et al. [96]. The study found out that 45 % of the respondents were willing

to pay $10/month, the other options were nothing (6 %), $5/month (21 %), $25/month (24 %) and $50/month (5 %). Education, age and income were the most determinant factors, but the significance was low. It was also noted that willingness to pay does not mean that the people would actually pay.

Faraqui et al. quantified customer response to dynamic pricing in California, USA. The study found that air conditioning ownership has a significant influence on demand response. Persons per household and electrical cooking on the other hand, have a negative correlation with demand response [97].

Baltimore Gas and Electric (U.S.) offered customers 10$/month during summer to be able to control air conditioning systems and residential water heaters via a VHF signal (unidirectional) [93]. The programme had 250 000 enrolled customers after 20 years.

Studies have been conducted in Europe as well. Pepermans estimated the willingness to accept power outages in Flanders [98]. Only a relatively small share of the population is willing to switch to a lower reliability level in exchange for a small decrease in the electricity bill. Households are willing to pay €30 per year on average to have outages during off-peak in stead of peak periods.

The responses to time of use tariffs are measured in various studies, the response itself varied from study to study [99]. 40 households, with a higher than average interest in electricity consumption, participated in a demand response programme in Norway [77]. The study found a 1 kWh/h demand response for customers with an electrical water heater [77].

A shift to off-peak periods was found in a Swedish programme with 50 participating households. The shifts increased at the end of the study, indicating a growing awareness of the participants [76].

The wind energy gains, financial savings and peak-time load reductions by directly controlling dishwashers on a fifteen minute scale have been tested in Ireland [100]. An optimisation algorithm shifted dishwashers in time. The minimise cost optimisation on market predictions resulted in 17.5 % financial savings and the highest reduction in peak demand. The wind generation demand increased by 26 % in the maximise wind optimisation on market predictions.

A field test in Italy found a higher average electricity demand and lower customer payments after implementing a time of use tariff scheme. The morning peak shifted in time, but the evening peaks remained [90].

Simulations based on Chinese data are conducted by He et al. [92]. The authors found a demand response of 8.41 % when prices are increased by 20 % and a response of 21.26 % if the increase is 40 %.

### 4.4.3 Attitude towards active demand

Stragier et al. determined the attitude of customers towards active demand with respect to appliances described in Section 2.4 [11]. In contrast to the weighting in Section 2.4 (age and inhabitants), the data is weighted according to the age of the respondent.

The basis for the attitudes is the Technology Acceptance Model [101, 102], using four dimensions to describe attitudes: perceived ease of use, perceived usefulness, attitude towards using and intention to use. Next to the four dimensions of the Technology Acceptance Model, five other dimensions are considered: safety, control, comfort, environment and price.

The attitude segments or groups are found by applying k-means clustering to the nine dimensions. The resulting attitude segments are

**advocates** having a positive perception of and a positive attitude towards smart appliances. (36 % of the weighted survey)

**supporters** having a more balanced though positive perception of smart appliances. They are sceptical about the level of control over the appliances and question safety. (27 % of the weighted survey)

**sceptics** having a sceptical perception of smart appliances: their attitude is only slightly negative. (25 % of the weighted survey)

**refusers** having a negative perception of and a negative attitude towards smart appliances. (12 % of the weighted survey)

### 4.4.4 Control strategies

Several approaches have already been proposed for demand response: central, decentral and hierarchical control strategies [103, 104].

The electricity demand of individual appliances are centrally scheduled in a centralised approach; in decentralised approaches, the appliances make their own decisions. A hierarchical approach combines both: optimisations are not done by one central system, but by multiple interacting central systems. Each appliance communicates with one of the central systems.

In the literature, the control strategies mainly focus on plug-in (hybrid) vehicles. An overview of control strategies for electric vehicles is given by Leemput et al. [103].

A comparison between a central (quadratic programming) and a hierarchical approach by Mets. et al [105], concluded that a central approach gets closer to the optimum. Atenzi et al. [106] compared optimisation on household level and global optimisation and noticed that the results of both approaches are equivalent.

### Centralised

Brooks et al. [13] described the use of electric vehicles and appliances for balancing. The TSO sends a dispatch signal, the aggregator gets the signal, decomposes it in order to control individual appliances or electric vehicles.

Logenthiran et al. [107] defined load shifting as an optimisation problem using a heuristic based genetic algorithm.

Chen et al. [108] compared stochastic optimisation to robust optimisation in a real time price scenario. Mixed integer linear programming is used for both optimisation approaches. The stochastic approach is computationally more expensive, but yields lower costs.

### Decentralised

Pipattanasomporn et al. [109] proposed a muti-agent system to control a microgrid using a control agent, an agent responsible for the distributed energy sources (DER), a user agent and a database agent.

Pedrasa et al. [110] used particle swarm optimisation algorithm to schedule appliances. The local optimisation is done centrally in home. However, the local optimisation can be seen as an agent in the electricity system as a whole.

A multi-agent system based load shedding algorithms is proposed by Xu et al. [111]. The agents are only able to communicate with their neighbours to decide which load to shed. However, an offline central particle swarm optimisation algorithm needs to determine the information to be communicated. The approach is hence only usable for a system that seldom suffer from power losses in the distribution lines.

### Hierarchical

Powermatcher [112] was the first market-based hierarchical multi-agent system to control devices. Local market agents send price signals to devices, able to

bid for the electrical energy they require. The local market agents communicate with a higher level market agent to distribute the prices.

Vandael et al. [104] combined distributed constraints of plug-in (hybrid) electrical vehicles with centralised optimisation. Vehicles send their constraints to a concentrator agent aggregating them. The concentrators send the aggregated constraints to an auctioneer agent, which minimises the overall cost. The electricity is distributed over the vehicles, based on their constrains, i.e. the price over time they are willing to pay.

The above approaches work with time slots. In [113], a comparison of working with intervals and operating in real time was made. Compromises had to be made in the real time system to reduce the communication. However, vehicles can be controlled faster. The method got improved later on with better optimisation algorithms, resulting in a better cost optimality [114].

## 4.5 Toolset

A software infrastructure is required for the data analysis. An overview presented in the literature is therefore given, together with an overview of security patterns, ensure confidentiality, integrity, availability and accountability.

### 4.5.1 Infrastructure

The work of Rusitschka et al. [115] proposed a cloud computing model able to manage real-time streams of smart grid data. The communication with the 'smart grid data cloud' is done via a web based application programming interface (API), more specifically representational state transfer (REST). 'Put' APIs allow sensors to place data into the cloud, while market actors are able to retrieve data using get APIs. The data management is done in a distributed way and data is processed in parallel. They refer to three approaches for protecting confidentiality: separate databases, separate schemas and shared schemas [116], but do not choose one. Pseudonymisation and aggregation are proposed as privacy protection measurements.

A data analysis architecture for power grids companies is proposed by Bâra et al. [117]. The grid database is coupled with a central data warehouse which supports online analytical processing (OLAP) and allows access for data mining techniques. A portal gives users access to the OLAP and the data mining frameworks.

Liu et al. [118] describe an infrastructure for grid models. Dispatchers, security engineers, planners, etc. are able to collaborate on modelling the grid.

An important factor is data analysis is data cleansing. Chean et al. [119] present techniques to handle locally and globally corrupted data by using smoothing techniques.

### 4.5.2 Security patterns

Security patterns are solutions to recurring information security problems [120]. Several security patterns exist in the literature. However, not all of them are well supported [121]. The most common security properties are: confidentiality, integrity, availability and accountability [122, 123].

Yoder et al. [124] already presented architectural patterns for application security in 1997. Seven security patterns are discussed in the work: single access point, check point, roles, session, full view with errors, limited view and secure access layer.

The security pattern repository of Kienzle et al. [120] describes two groups of security patterns. Structural patterns are comparable to design patterns [125]. They provide a simple and elegant solutions to security problems in the form of diagrams of structure and descriptions of interactions. Procedural patterns on the other hand are designed to improve the process of developing secure software.

A more extensive overview of security patterns is given by Yskout et al. [126]. Three categories are distinguished: application architecture, application design and system. 35 security patterns are described in total.

Heyman et al. [121] give an overview of the number of security patterns, the domains they operate in and the quality of the patterns. They state that there are too many patterns, whilst not every pattern fits the definition of a security pattern. The quality of the documentation of security patterns is often lacklustre and privacy and non-repudiation aren't supported well by security patterns.

Yoshioka et al. [122] gave a survey on security patterns. Security patterns are categorised according to software development phases: the requirement phase, the design phase and the implementation phase. The ease of use, the effectiveness and the sufficiency of security patterns of patterns are discussed as well.

Yautsiukhin et al. [123] defined a quantitative metric for security patterns

to measure how well an architecture is protected against relevant security threats. The severity *sev* of a threat is computed as the normalised sum or the reproducibility, the exploitability and the discoverability. The threat protection is calculated as one minus the multiplication of the coverage of the implemented patterns. The protection against an attack is the summation over all threats of the multiplication of the severity of the threat by the protection against the threat.

## 4.6 Conclusions

The literature related is presented in this chapter. Topics are description of electricity demand of households and appliances; the possibilities to detect appliances in load profiles and the potential privacy issues; ways to model electricity demand; driving factors for demand response, studies and field tests related to demand response, the attitude of customers towards active demand and control strategies for active demand; and software to perform data analyses on electrical load data.

# Chapter 5

# Electricity demand description

The development of smart grid integration strategies requires knowledge about electricity demand. Aggregated demand as total electricity demand [127] and average load profiles [128] is sufficient when focus is on global results. Detailed electricity demand profiles are required for the simulation of voltage problems caused by vehicle charging [129] and micro-grids [130] or to estimate the potential of battery storage in a distribution grid [131]. Two ways exist to deliver input for simulations: data selection and data generation.

Data selection requires a thorough description of demand, both in terms of the distribution of annual electricity demand as well as demographic parameters related. Annual demand only expresses the total demand, while demographic properties have an influence on the timing of demand and allow for the construction of districts. The demographic properties found are compared to those described in the literature.

Customers are clustered into groups based on timing and magnitude of demand to allow for load profile generation (Section 7.2). The number of groups is limited to double the number of customer types defined by the regulator, to reduce the risk of outliers. The group or cluster centre found after analysis, represents the average consumption pattern of similar customers and allows for data up-scaling.

Electricity demand of wet appliances is described according to the found groups. The measurement data of Linear is scaled up for the description, given the limited number of households with measurements at both connection point and of appliances. The scaling up process consists of spreading the data over the customer groups.

## 5.1 Demographic description

Knowledge of the total annual electricity demand and the correlated demographic parameters make it easier to understand trend and volume patterns, not being the focus of this section but described in Sections 5.2 and 5.3 and Chapters 7 and 8.

The distribution of the total annual electricity demand for Flanders is explained in Section 5.1.1. Customer types described by the Flemish regulator for Electricity and Gas (VREG) are listed and with measurement data from distribution network operators. Probability density functions are to parametrise the histogram.

An expert is needed to find patterns in large amounts of data. Machine learning algorithms are expert systems which are able to decide which parameters are relevant. Here, they are used to determine the demographic parameters for Belgium/Flanders from responses to a questionnaire and measurement data from distribution system operators. The common supervised machine learning algorithms are explained in Section 3.2.1.

The algorithms are applied to measurement data and questionnaire to find the demographic parameters related to the total electricity demand (Section 5.1.2). Evaluation criteria to find the best performing machine learning algorithm are explained as well. Different demographic parameters are the input for the different machine learning algorithms. The input parameters for the best performing algorithms are regarded the relevant demographic parameters. The algorithm with the highest receiver operating curve area under curve, true positive rate and precision and with the lowest false positive rate for each class to predict is regarded best performing.

### 5.1.1 Total annual electricity demand

The Flemish Regulator for Electricity and Gas (VREG) describes consumer types according to their total annual electricity demand (Table 5.1) [132]. The corresponding relative number of households for each type is derived from measurement data from distribution network operators and is validated (by comparing distributions) against a retailer's customer data [133].

A more detailed view on the total annual electricity demand is obtained by presenting the measurement data in a histogram. Figure 5.1 depicts the distribution of the total yearly electricity demand. The mean of the distribution is 4.9 MWh, the mode is 2.8 MWh and the median is 3.6 MWh.

Table 5.1: Consumer types according to the Flemish regulator

| Type | Total annual electricity demand [kWh] | Households [%] |
|---|---|---|
| Small | $d < 900$ | 4.7 |
| Relatively small | $900 \leq d < 2350$ | 18.8 |
| Average | $2350 \leq d < 5500$ | 47.5 |
| Relatively large | $5500 \leq d < 13750$ | 23.5 |
| Large | $13750 \leq d$ | 5.5 |

Three curves are fitted through the histogram to parametrise the distribution using a non-linear least-squares solver. Only skewed functions were considered because of the skewness of the distribution. The tested probability density functions with their respective parameters are Weibull (5.1), Rayleigh (5.2) and Log-normal (5.3) [134].

$$pdf_{weibull} = k \cdot \frac{x^{k-1}}{\lambda^k} \cdot e^{-\left(\frac{x}{\lambda}\right)^k} \tag{5.1}$$

$$with \quad \left| \begin{array}{l} k = 1.7831 \\ \lambda = 4.4418 \end{array} \right.$$

$$pdf_{rayleigh} = \frac{x}{\sigma^2} \cdot e^{-\frac{x^2}{2\sigma^2}} \tag{5.2}$$

$$with \quad \left| \begin{array}{l} \sigma = 3.0751 \end{array} \right.$$

$$pdf_{log-normal} = \frac{1}{x \cdot \sigma \cdot \sqrt{2\pi}} \cdot e^{\frac{-(\ln x - \mu)}{2\sigma^2}} \tag{5.3}$$

$$with \quad \left| \begin{array}{l} \mu = 1.3604 \\ \sigma = 0.6854 \end{array} \right.$$

To verify which distribution fits the data best, the goodness of fit is tested with the Pearson's Chi Squared test [135]. The $\chi^2$-test prescribes that a fit is accepted if the $\chi^2$-error is lower than the critical value of the required confidence level. The critical value itself depends on the degrees of freedom for the fit. The degrees of freedom are determined by the number of classes, i.e. the number of histogram bars with more than five elements, and the number of parameters

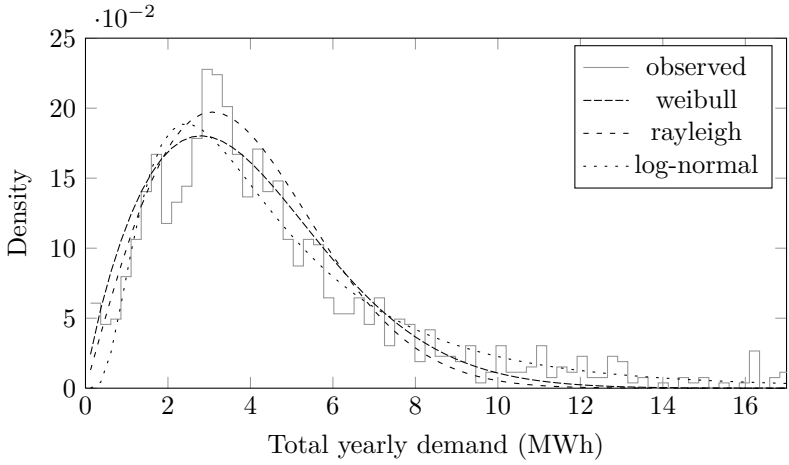



Figure 5.1: Distribution of total yearly electricity demand and curve fits

of the probability density function. The critical values based on the degrees of freedom and the confidence level are described in tables in books [135].

The $\chi^2$-error is the squared sum of the difference between observed ($O$) and expected ($E$) outcome frequencies divided by the expected frequencies as for all classes with more than five elements ($N$).

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - Ex_i)^2}{Ex_i} \tag{5.4}$$

The first five element class is the $38^{th}$ class, meaning that the last non five element class is the $37^{th}$. The Log-normal function has two parameters, the degrees of freedom are $(37-2) = 35$. Rayleigh has only one parameter, resulting in 36 degrees of freedom. An overview of the $\chi^2$-errors is to be found in Table 5.2. Both Weibull and Rayleigh are accepted as fits, even with a confidence level of 99 %. Log-normal on the other hand is rejected because the $\chi^2$-error is much larger than the critical value of the 95 % confidence level. The Weibull distribution fits the data slightly better than Rayleigh based on the $\chi^2$-error. However, the Weibull distribution underestimates the number of high consuming customers (Figure 5.1).

Table 5.2: Pearson $\chi^2$ test results

| Distribution | $\chi^2$ | Critical value for confidence of | |
|---|---|---|---|
| | | 95 % | 99 % |
| Weibull | 0.24 | 22.46 | 18.51 |
| Rayleigh | 0.41 | 23.27 | 19.23 |
| Log-normal | 259.8 | 22.46 | 18.51 |

### 5.1.2 Mapping properties to electricity demand

To find the demographic properties related to the total annual electricity demand, a function is mapped between various demographic parameters and the total annual electricity demand. The performance of the function depends on the relevance of the input parameters with respect to the output. To determine the best function, multiple combinations of input parameters are tried with various functions. The better the function is at predicting the outcome, the more relevant the input parameters are.

#### Data selection and preparation

Two data sources are used: the questionnaire described (Section 2.4) and the corresponding measurement data (Section 2.3). The answers to the questionnaire represent the demographic properties of the respondent. Each question is thus an input parameter, the electricity demand is the output parameter. Only the respondents with measurement data of 2008 (416 in total) are considered.

#### Data transformation

The answers of the respondents to the various questions need to be converted into a useful form for the machine learning algorithms. The machine learning algorithms (Section 3.2.1) work with numerical values as input. Hence, the answers to the questions need to be converted in numbers. This translation in natural numbers ensures order in the answers, for example a surface area range of 0 to 50 $m^2$ gets number 1, 50 $m^2$ to 99 $m^2$ gets number 2, etc. The number of inhabitants is already a natural number and doesn't need conversion. Other translations are less straight forward, e.g. there is no order in heating sources (gas, electricity, fuel oil, butane).

The use of natural numbers poses no problems for decision trees. Decision trees accustom their node tests to work around differences in inputs. However,

support vector machines and neural networks require normalisation: the input has a value between and 0 and 1.

The outcome needs to be transformed as well. A floating point number is harder to predict with machine learning techniques than classes. The total annual electricity demand is made discrete according to the customer type of the Flemish regulator (Table 5.1). However, the class 'average' consumer consists of 47.5% of the data. Small classes as 'small' and 'large' consumer are regarded as outlier, grouped together with respectively 'relatively small' and 'relatively large' resulting in three classes: 'small', 'average' and 'large'.

### Mining

Machine learning techniques have their own way of fitting a function mapping input to output. The best performing technique of creating a function that maps input to output is preferred. Various machine learning techniques are therefore compared in the mining step of the KDD process.

Only supervised machine learning algorithms are selected for data mining as inputs and output are known. Three supervised learning algorithms are compared: Sequential Minimal Optimization (Support Vector Machine technique), Multilayer Perceptron (a Neural Network approach), and C4.5 (a Decision Tree algorithm) (Section 3.2.1). To get the maximum out of the data set, the algorithms are executed with ten-fold cross validation. For ten-fold cross validation, data is divided randomly in ten different equally large sets. The machine learning algorithm is executed with nine of the ten sets, the remaining set is used to calculate the error rate. This is performed ten time so that each set is used once as validation set. The average of the ten error estimates is the overall error.

In order to find the best input parameters for the output wanted, different inputs are combined. Questions of the questionnaire are added and/or removed, based on the evaluation of the fitted function.

### Interpretation

Sensitivity, false positive rates, precision and receiver operating characteristic (ROC) area under the curve are used to verify how well each algorithm performs. The basis for these metrics is the confusion matrix (Table 5.3). The columns of a confusion matrix capture the actual class of the instances, which can be positive ($P$) or negative ($N$). The rows indicate the predicted class and are marked with a prime. The true positives ($TP$) in the confusion matrix is the number

| | P | N |
|---|---|---|
| P' | TP | FP |
| N' | FN | TN |

Table 5.3: Confusion matrix

of instances correctly classified positive. The other elements of the confusion matrix are false positives ($FP$), false negatives ($FN$) and true negatives ($TN$).

Sensitivity is the true positive rate ($TPR$) and expresses to which extend an instance is correctly considered positive (5.5). The false positive rate ($FPR$) on the other hand shows the degree at which a negative is classified positive (5.6). Precision or positive predicted value ($PPV$) is the rate of positively classified instances that are actually positive (5.7).

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{5.5}$$

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP} \tag{5.6}$$

$$PPV = \frac{TP}{P'} = \frac{TP}{TP + FP} \tag{5.7}$$

Sensitivity and false positive rate are picked because these metrics also exist in statistics. A type I error is a false positive error, while type II errors are false negative errors, the rate is expressed as $1 - TPR$.

A receiver operating curve depicts the performance of a binary classifier by plotting the true against the false positive rate. For each positively classified instance, the correctness of the prediction is checked. The curve goes up one step, if the prediction is correct and to the right by one step if it is not. The more the curve lays in the upper left corner, the better the algorithm performs. As ROC curves take a lot of space, the area under the curve (AUC) is calculated, representing the curve by one number. A perfect classifier has a ROC AUC of 1. An ROC AUC of 0.5 indicates predicting as good as a random model.

The goal of the algorithms is to predict each class as well as possible. The ROC AUC, TPR and PVV need to be as high as possible for each class, while the FPR needs to be as low as possible. ROC AUC, TPR and PVV values are considered to be good if they are greater than 0.7, FPR values need to be lower than 0.3.

Table 5.4: Evaluation of machine learning algorithms.

| | TPR | FPR | Precision | ROC AUC | Class |
|---|---|---|---|---|---|
| SMO | 0.067 | 0.000 | 1.000 | 0.708 | small |
| | 0.937 | 0.791 | 0.573 | 0.574 | average |
| | 0.298 | 0.047 | 0.720 | 0.684 | large |
| MLP | 0.253 | 0.079 | 0.413 | 0.788 | small |
| | 0.806 | 0.653 | 0.583 | 0.606 | average |
| | 0.364 | 0.071 | 0.677 | 0.722 | large |
| *C4.5* | *0.307* | 0.038 | 0.639 | 0.767 | small |
| | *0.829* | *0.607* | *0.607* | *0.611* | average |
| | *0.438* | 0.088 | 0.671 | 0.708 | large |

In initial test, it was clear the the ROC AUC for each class was lowest for SMO, the ROC AUC of MLP and C4.5 were comparable. C4.5 performed best because the type I ($FPR$) and the type II ($1 - TPR$) errors and precision ($PVV$) were lower compared for each class to SMO.

To compare the three algorithms, the best input parameters of C4.5 are also tested with SMO and MLP. The difference in performance of the algorithms is presented in Table 5.4. SMO with a second order polynomial kernel shows the worst performance: the algorithm labels most instances as 'average', little as 'large' and almost none as 'small'. MLP performs better, more instances are labelled 'small' and 'large'. However, C4.5 has a larger true positive rate, a better precision and a similar ROC AUC. The results in italic show where C4.5 performs best compared to the other techniques: 6 of the 12 criteria.

When using normalized units, the area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). This can be seen as follows: the area under the curve is given by (the integral boundaries are reversed as large T has a lower value on the x-axis)

### Knowledge

The most discriminating parameters according to the C4.5 model are

- age of the respondent,
- number of inhabitants,
- number of inhabitants at home during the evening,

- surface area for a business,
- owning a freezer,
- owning a dishwasher,
- housing type.

No order could be determined in the parameters. The results are similar to the literature (Section 4.1.1).

## 5.2 Groups based on electricity demand

Demographic properties only describe electricity demand partially (Section 4.1.1). Quantifying similarities amongst demand of customers is another way of defining electricity demand. As described in the literature (Section 4.1.1), statistical and cluster analysis are capable of finding similarities.

The objective is to find groups of customers with a similar electricity demand pattern. As for the SLP profiles, day and night consumers have to be found. The inclusion of business consumers is regarded as a plus. The number of customer groups should be limited to 10, double the number of of customer types defined by the regulator (Table 5.1), limiting the number of outlier groups. The group or cluster centre found after analysis, needs to represent the average consumption pattern of similar customers and need to allow for data upscaling.

### 5.2.1 Data preparation

The input data for the clustering analysis is the residential measurement data delivered by the distribution network operators (Section 2.3). The data transformation is the conversion of the load profiles into load curves (Section 2.1).

#### Selection and pre-processing

The distribution network operators ensure that measurements are performed on a representative sample of the Flemish population by means of random sampling. From those measurements, the active power for one year (2008) is selected, resulting in a load profile for each customer. The distribution network operators cleaned the data before delivery.

### Transformation

The load profiles consist of fifteen minute measurements for one year, this results in 35040 measurement points for a normal and 35136 in a leap year. The number of points are regarded dimensions in the proposed clustering algorithms. More dimensions result in more difficulty for the clustering algorithm to reject an instance from the cluster and hence a higher computational cost [136]. Moreover, the clustering algorithm could find more clusters by finding similarities in less relevant dimensions.

Various electricity load profile transformation techniques are described in the literature. Amongst them are Harmonics [44], Principal Component Analysis [137], normalized daily load patterns [46] and normalized load curves [38]. The normalization in [46] consists of dividing each dimension of a load profile by the maximum value in that load profile. The approach has a big disadvantage: the magnitude of consumption is lost, only variations in electricity consumption are considered. Z-score normalisation (the number of standard deviations an observation is above the mean) is used in [38].

The load profiles are converted into load curves as a dimension reduction technique. A load curve of the average days of the week for the different quarters of the year, i.e. the average power every fifteen minutes of the day, for every day of the week, for an average week every quarter, consist out of 2688 dimensions, lower than the initial 35040. Weekdays are correlated as indicated by autocorrelation plots (Figure 7.3). However, the autocorrelation might not be true for all households and information about holidays is lost when applying the data dimension technique.

The load curves are not normalized for the EM-clustering (observations are projected on Gaussian functions, which standardises the errors in clustering), but normalized for the KM clustering within the EM-clustering algorithm.

Min-max normalization scales the values in each dimension, the normalized value in dimension $d$ of vector $i$ is represented by $a_{i,d}$.

$$a_{i,d} = \frac{x_{i,d} - \min x_{i,d}}{\max x_{i,d} - \min x_{i,d}} \tag{5.8}$$

The value of instance $i$ in dimension $d$, i.e., $x_{i,d}$ is reduced with the minimal value over all instances in that dimension and divided by the range between the minimal and the maximal value over all instances in that dimension [4]. For the KM clustering algorithm, min-max normalisation is used. Min-max projects values linearly onto a value between 0 and 1. Z-score on the other hand, projects the data around 0 and mostly not in a range between 0 and 1.

### 5.2.2 Pattern detection and interpretation

The applied mining algorithm is Expectation Maximisation clustering. The algorithm found 10 clusters, containing customers who mainly consume during the day, the night and with a business at home.

#### Mining

A clustering technique with weights is preferred, used to scale the data up (Section 5.2.3). In Section 4.1.1, load profile and load curve clustering techniques from the literature are described. The clustering techniques involving weights are FKM and EM clustering and SOM. SOM is not considered, because it had the worst performance in other work (Section 4.1.1). FKM clustering is a centroid based technique, while EM clustering is density based. Distribution based techniques model the data in distributions, making the theoretical foundations more sound. better at handling randomness in data. The technique also captures correlation and dependence of attributes. Therefore, EM clustering is selected as clustering algorithm.

According to the literature, KM and EM clustering are the amongst the best algorithms to cluster load curves: they have the lowest cluster dispersion. EM clustering incorporates KM clustering in its initialization step. After initialization, a normal distribution of each dimension of the data is calculated for each cluster. The distributions are updated until the overall likelihoods stop changing significantly. The complete algorithm and the KM clustering algorithm are both explained in Section 3.2.2.

#### Groups of customers

Expectation maximisation clustering is applied to customers' load curves and found 10 clusters with corresponding centre mean and standard deviation per dimension. The different clusters are named afterwards according to the total electricity demand on an annual basis, the timing of electricity demand during the day and the possibility of having a business at home (Table 5.5). The latter is derived by linking answers to questionnaires (Section 2.4) to the clusters.

The percentage of customers in a cluster is the prior probability of the cluster $P(S_{cl})$ (3.20). The estimated total annual electricity demand of a cluster $E_{cl,est}$ is derived from the average power values of the cluster centre, the number of

Table 5.5: Groups of customers

| Type | Sub-type | Total yearly [kWh] | prior [%] |
|---|---|---|---|
| Day | small | 800 | 14.1 |
| | relatively small | 2 500 | 25.9 |
| | average | 4 250 | 27.8 |
| | relatively large | 6 650 | 15.4 |
| | large | 11 600 | 7.7 |
| Night | average | 6 200 | 3.2 |
| | large | 8 750 | 2.9 |
| Business | average | 28 350 | 2.3 |
| | relatively large | 70 000 | 0.5 |
| | large | 189 600 | 0.1 |

days in a year and the number of hours in a day,

$$E_{cl,est} = \frac{\sum_{d=1}^{n_d} m_{cl,d}}{n_d} \cdot 365 \cdot 24 \tag{5.9}$$

The most important clusters are these with the highest probability. Relatively large and large business are considered to be outliers. The consumers are more spread over the customer groups compared to the customer types of the regulator (Table 5.1): there are no groups with more than 30% of the consumers and within the day consumer groups, there are no groups representing less than 5%.

### 5.2.3 Clusters as weighting model

The number of dimensions after the data transformation remains high ($n_d = 2688$). The probability that an instance $i$ belongs to a cluster is influenced by every Gaussian distribution around the cluster centre: the probability density values are multiplied with each other and with the probability of the cluster (the prior). An instance closest to a cluster centre, is included in the cluster centre with a high probability because of the multiplication of the probability density values. However, each dimension adds to the cluster representation, e.g. the difference between day and night consumers.

The reduction of the number of dimensions relaxes the attraction of the instances towards the closest cluster centre. Two different relaxations are considered: reducing the dimensionality of both likelihoods and priors or reducing only the

dimensionality of the likelihoods. The former corresponds to the assumption that the model is correct and checks the probability that the instance was created by the model. The latter corrects the previous assumption by adding information which states that some clusters are more probable than others.

### Relaxed model

The reduction of the dimensionality of both the likelihoods and the priors is done by taking the $(n_d + 1)$'th root of the multiplication of both. A numerical more stable method which accomplishes the same is dividing the logdensity of Equation 3.26 instead,

$$\log dens_{i,cl} = \frac{\log P(S_{cl}) + \sum_{d=1}^{n_d} \log pdf_{i,cl,d}}{n_d + 1} \tag{5.10}$$

The probability of a cluster $P(S_{cl})$ is negligible compared to the probability density function values *pdf* because of the high value for $n_d$ (5.10). The equation can thus be interpreted as the average of the likelihoods,

$$P(\mathbf{x}_i|S_{cl}) \approx \exp\left(\log dens_{i,cl}\right) \approx \exp\left(\frac{\sum_{d=1}^{n_d} \log pdf_{i,cl,d}}{n_d}\right) = P_r(\mathbf{x}_i|S_{cl}) \tag{5.11}$$

The probability of an instance belonging to a cluster is hence approximately the normalized likelihood that the instance is drawn from the considered cluster,

$$P_r(\mathbf{x}_i \in S_{cl}) = \frac{P_r(\mathbf{x}_i|S_{cl})}{\sum_{j=1}^{n_{cl}} P_r(\mathbf{x}_i|S_j)} \tag{5.12}$$

The denominator needs to be constant to have normalisation, which means that in (5.12), it is indirectly assumed that the probability of each cluster is equal. The idea behind the assumption is that the cluster models are assumed to be correct when building with them. The result of the relaxation is that data from clusters with high probability will enter clusters with a lower probability. Clusters with higher demand for electricity have their average demand lowered, while clusters with a lower demand for electricity see the opposite. Extremes are pulled more towards the mean.

### Corrected relaxed model

The spread of instances over different clusters can be damped by compensating the equal probability of clusters. The corrected relaxation is accomplished by, instead of projecting both likelihood and priors to one dimension, projecting only the likelihood on one dimension: averaging the loglikelihood only over the Gaussians,

$$\log dens_{i,cl} = \frac{\sum_{d=1}^{n_d} \log pdf_{i,cl,d}}{n_d} + \log P(S_{cl}) \tag{5.13}$$

The likelihood of being drawn from cluster $S_c$ becomes,

$$Pr(\mathbf{x}_i|S_{cl}) = \exp\left(\frac{\sum_{d=1}^{n_d} \log pdf_{i,cl,d}}{n_d}\right) \tag{5.14}$$

The probability of an instance belonging to a cluster is hence the original Bayes' theorem where the likelihood of being drawn from a cluster $P(\mathbf{x}_i|S_{cl})$ is replaced by the $n_d^{th}$ root of that likelihood.

$$P_{cl}(\mathbf{x}_i \in S_{cl}) \approx \frac{P(S_{cl}) \cdot Pr(\mathbf{x}_i|S_{cl})}{\sum_{j=1}^{n_{cl}} P(S_j) \cdot Pr(\mathbf{x}_i|S_j)} \tag{5.15}$$

### Trade-offs

In each of the three approaches, for each instance, the most probable cluster to belong to is the same. If the original cluster membership approach indicates that e.g. the average day consumer cluster is the most probable for the instance, the cluster is most probable in the other approaches as well. The difference in the approaches is the magnitude of the probability.

The sum over all instances of the probabilities belonging to a given cluster represents the probability of the cluster (the prior). The priors for every cluster are calculated for the original cluster membership, the relaxed cluster membership and the corrected relaxed cluster membership (Table 5.6).

The relaxation spreads the membership of one cluster over multiple clusters, in general the neighbouring clusters, as can be derived from the correlation plot over the different types of customers (Figure 5.2 (a)). The 'd' in the correlation

Table 5.6: Prior cluster probabilities for the different cluster membership approaches

| Type | Sub-type | Prior [%] | | |
|---|---|---|---|---|
| | | Original | Relaxed | Corrected |
| Day | small | 14.1 | 8.8 | 7.7 |
| | relatively small | 25.9 | 16.8 | 22.3 |
| | average | 27.8 | 20.6 | 31.4 |
| | relatively large | 15.4 | 18.6 | 19.3 |
| | large | 7.7 | 13.3 | 9.9 |
| Night | average | 3.2 | 3.1 | 2.3 |
| | large | 2.9 | 10.5 | 3.3 |
| Business | average | 2.3 | 6.7 | 3.2 |
| | relatively large | 0.5 | 1.5 | 0.5 |
| | large | 0.1 | 0.2 | 0.1 |

plot stands for 'day', 'n' for 'night' and 'b' for business. The subtypes are 'small' (s), 'relatively small' (rs), 'average' (a), 'relatively large' (rl) and 'large' (l). Because of the equal probability assumption, a threshold can be placed to find the most corresponding cluster. When a threshold is placed at a probability of 1/3, 86.94 % of the instances (1185) have the same cluster as the original, 7.12 % (97) have a double label and 5.94 % (81) have none.

The corrected relaxation corrects the spread of the cluster membership and makes the correlation between clusters with a high probability weaker (Figure 5.2 (b)). The correlation is higher in the clusters with a low probability because of the correlation between the low numbers.

The electrical power distribution of the different approaches are visualised in Figure 5.3. The middle bar represents the power distribution in the original case, left is the relaxed and right the corrected relaxed model. Minimum and maximum value are the $5^{th}$ and the $95^{th}$ percentile, the box boundaries are placed at the $25^{th}$ and the $75^{th}$ percentile and the line in the box represents the median or the $50^{th}$ percentile.

The spread of the cluster membership due to the relaxation results in higher electrical powers for the clusters with low and lower electrical powers for the clusters with high electrical power. Both high and low powers regress to the mean. The correction on the relaxation limits this regression, but does not fully compensate for it.

| | d-s | d-rs | d-a | d-rl | d-l | n-a | n-l | b-a | b-rl | b-l |
|---|---|---|---|---|---|---|---|---|---|---|
| d-s | 1.00 | 0.44 | -0.41 | -0.77 | -0.55 | 0.22 | -0.02 | -0.32 | -0.10 | 0.04 |
| d-rs | 0.44 | 1.00 | 0.42 | -0.58 | -0.82 | -0.20 | 0.04 | -0.68 | -0.38 | -0.07 |
| d-a | -0.41 | 0.42 | 1.00 | 0.40 | -0.36 | -0.49 | 0.00 | -0.60 | -0.49 | -0.15 |
| d-rl | -0.77 | -0.58 | 0.40 | 1.00 | 0.61 | -0.39 | -0.16 | 0.12 | -0.15 | -0.10 |
| d-l | -0.55 | -0.82 | -0.36 | 0.61 | 1.00 | -0.14 | -0.31 | 0.73 | 0.30 | 0.01 |
| n-a | 0.22 | -0.20 | -0.49 | -0.39 | -0.14 | 1.00 | 0.27 | 0.07 | 0.14 | 0.12 |
| n-l | -0.02 | 0.04 | 0.00 | -0.16 | -0.31 | 0.27 | 1.00 | -0.21 | -0.10 | 0.01 |
| b-a | -0.32 | -0.68 | -0.60 | 0.12 | 0.73 | 0.07 | -0.21 | 1.00 | 0.74 | 0.09 |
| b-rl | -0.10 | -0.38 | -0.49 | -0.15 | 0.30 | 0.14 | -0.10 | 0.74 | 1.00 | 0.14 |
| b-l | 0.04 | -0.07 | -0.15 | -0.10 | 0.01 | 0.12 | 0.01 | 0.09 | 0.14 | 1.00 |

(a) relaxed

| | d-s | d-rs | d-a | d-rl | d-l | n-a | n-l | b-a | b-rl | b-l |
|---|---|---|---|---|---|---|---|---|---|---|
| d-s | 1.00 | 0.49 | -0.42 | -0.67 | -0.32 | 0.11 | 0.00 | -0.05 | 0.05 | 0.08 |
| d-rs | 0.49 | 1.00 | 0.26 | -0.82 | -0.76 | -0.25 | -0.32 | -0.43 | -0.23 | -0.10 |
| d-a | -0.42 | 0.26 | 1.00 | 0.09 | -0.60 | -0.47 | -0.46 | -0.61 | -0.43 | -0.23 |
| d-rl | -0.67 | -0.82 | 0.09 | 1.00 | 0.56 | -0.12 | 0.00 | 0.02 | -0.10 | -0.05 |
| d-l | -0.32 | -0.76 | -0.60 | 0.56 | 1.00 | 0.16 | 0.23 | 0.54 | 0.23 | 0.11 |
| n-a | 0.11 | -0.25 | -0.47 | -0.12 | 0.16 | 1.00 | 0.55 | 0.29 | 0.26 | 0.19 |
| n-l | 0.00 | -0.32 | -0.46 | 0.00 | 0.23 | 0.55 | 1.00 | 0.35 | 0.28 | 0.18 |
| b-a | -0.05 | -0.43 | -0.61 | 0.02 | 0.54 | 0.29 | 0.35 | 1.00 | 0.67 | 0.19 |
| b-rl | 0.05 | -0.23 | -0.43 | -0.10 | 0.23 | 0.26 | 0.28 | 0.67 | 1.00 | 0.21 |
| b-l | 0.08 | -0.10 | -0.23 | -0.05 | 0.11 | 0.19 | 0.18 | 0.19 | 0.21 | 1.00 |

(b) corrected

Figure 5.2: Correlation plot customer types for relaxed and corrected model

## Conclusions

The most theoretically sound clustering algorithm described in the literature, together with the most common data reduction technique (Section 4.1.1) has been applied to the load profiles of 1363 customers (Section 2.3).

The clustering algorithm was able to distinguish day and night consumers and also business consumers were found. The number of clusters allowed has been limited to ten, which is also the number of clusters found. More clusters results in a higher overall likelihood, but increases the number of outlier clusters. Large business consumers is considered to be an outlier, as well as the relatively large business consumer but to a lower extend. The inclusion of both in the average business consumer would have a large impact on it, given the high electricity demand of them.

The customers are more spread over the customer groups compared to the

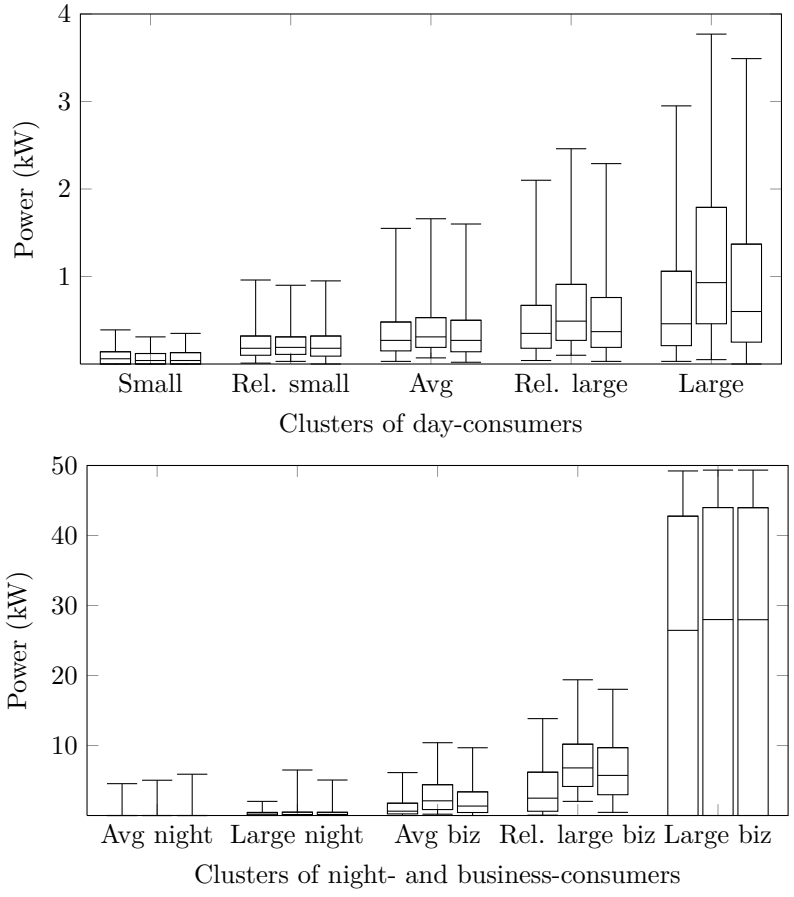



Figure 5.3: Distribution of electrical power for original (centre), relaxed (left) and corrected (right) cluster data.

consumer types of the regulator (Table 5.1): the regulator's 'small consumer' only represents 4.7% of the customers while 'the small day consumer' here covers 14.1%, also no customer group represents more than 30%.

The cluster centre is a load curve consisting of the weighted average values per dimension of the load curves with standard deviation. The load curve represents the average days of the week per quarter with a resolution of fifteen minutes and is the average electricity demand within the customer group.

## 5.3 Appliance electricity demand description based on group membership

Electricity demand at the connection point of a household is composed of the demand of the different appliances. The decomposition of the electricity demand improves the understanding of electricity demand patterns. The focus is on appliances with a high power demand, which can easily be controlled for active demand: only wet appliances (washing machines, tumble dryers and dishwashers) are considered.

Measurements at the connection point of a household, are translated into the probability that the household is part of a customer type. The relaxed cluster membership algorithm of Section 5.2.3 is applied to calculate the probability, used to weight the load curve of the appliances of the household in the appliance load curve of the cluster.

### 5.3.1 Data preparation

Appliance measurement data and the corresponding measurements at the connection point from the Linear project (Section 2.5.1) are used to describe electricity demand by appliances. The number of households with both appliance measurements and measurements at the connection point is limited: 30 households with a washing machine, 27 with a tumble dryer and 21 with a dishwasher. Most of those households have an electricity demand varying between average relatively large. The number of households per customer group is limited if crisp cluster memberships are applied. To spread the measurements over the various customer groups, fuzzy cluster memberships are used, resulting in data up-scaling.

The relaxed clustering model of Section 5.2.3 (5.12) has to be applied to scale-up data. Therefore, the load profiles at the connection point, as well as the load profiles of the appliances, are converted into load curves.

#### Selection and pre-processing

The first requirement for the measurement data is having both appliance measurements and measurements at the connection point. Without measurements at the connection point, it isn't possible to determine the probability of being part of a cluster.

The second requirement is the reliability of the active power measurements. Measurements at the connection point are not always reliable (Section 2.5.1): only 42 of the 56 households have reliable measurements. The reliability criteria for a measurement at the connection point is a power of minimal 8 W. If the power is lower, the measurement is discarded because the point could not have been measured: the power demand of the measurement equipment is more than 8 W. For household to be included, at least a half year of reliable measurement points is required. The measurements of the appliances are more reliable because of the buffering properties of the measuring plugs.

#### Transformation

The load profiles at connection point and of the appliances are converted into load curves, representing the electricity demand during average days of the week for the different quarters of the year. The data transformation is the same as for the clustering (Section 5.2.3) because the clustering model is used to estimate the probability of households belonging to a customer type. Load profiles which cannot be converted into load curves are omitted.

### 5.3.2 Pattern detection and interpretation

The load curves at the connection point are translated into the relaxed probability of belonging to the various customer groups. The data of one household is hence spread amongst difference customer groups, which results in more data per customer group.

The customer group load curves of the wet appliances are constructed by combining the household's probability of belonging to the customer group with the appliance load curve of the household.

Finally, the resulting wet appliance load curves are compared to the literature values for validation purposes.

### Scaling data up

The electricity demand of an appliance for a customer group is derived from the appliance electricity demand of multiple households. The probability that a household is part of a given customer group (or cluster) determines the weight that a load curve of the appliance has in the load curve of the customer group.

The relaxed cluster model of Section 5.2.3 calculates the probability that a customer is part of a given cluster or customer group $P(\mathbf{x}_i \in S_{cl})$. The load curve of the appliance $\mathbf{x}_i^a$ gets weighted by the probability. The sum of the weighted load curves is normalized to obtain the load curve of the customer group $\mathbf{x}_{cl}^{app}$,

$$\mathbf{x}_{cl}^{app} = \frac{\sum_{i=1}^{n_i} P(\mathbf{x}_i \in S_{cl}) \cdot \mathbf{x}_i^{app}}{\sum_{i=1}^{n_i} P(\mathbf{x}_i \in S_{cl})} \tag{5.16}$$

A household with a probability of belonging to a customer group of more than 0.2 is considered to have a large weight in the customer group. Table 5.7 shows the sum of the probabilities given the customer group ($\sum_{i=1}^{n_i} w_{i,cl}$) and the number of probabilities higher than 0.2 for that group ($\#w_{i,cl} > 0.2$).

Various customer groups are under-represented as a result of the lack of customers with a (heigh) weight in the group. Insufficient information is available for small day-, average night-, relatively large business- and large business-consumers. No load curves are built for those customer groups.

### Appliance load demand

The appliance load curves are smoothed to reduce outlier impact. A data point of the smoothed appliance load curve is calculated by taking the average of previous, current and next data point of the original appliance load curve.

The mean, median and maximum values of the smoothed load curves of the various wet appliances for each customer type are shown in Table 5.8. The mean is an approximation for the total electricity consumption on a yearly basis. The median divides the distribution of the power of the load curve in two: power is

Table 5.7: Wet appliances: availability in Linear measurement data

| Type | Sub-type | Washing machine | | Tumble dryer | | Dishwasher | |
|---|---|---|---|---|---|---|---|
| | | $\sum_{i=1}^{n_i} w_{i,cl}$ | $\#w_{i,cl} > 0.2$ | $\sum_{i=1}^{n_i} w_{i,cl}$ | $\#w_{i,cl} > 0.2$ | $\sum_{i=1}^{n_i} w_{i,cl}$ | $\#w_{i,cl} > 0.2$ |
| Day | small | 0.02 | 0 | 0.02 | 0 | 0.02 | 0 |
| | relatively small | 2.26 | 6 | 1.92 | 5 | 1.40 | 4 |
| | average | 8.27 | 23 | 6.93 | 19 | 5.50 | 16 |
| | relatively large | 9.33 | 24 | 8.61 | 22 | 6.45 | 17 |
| | large | 5.38 | 9 | 5.14 | 10 | 3.70 | 7 |
| Night | average | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 |
| | large | 2.78 | 2 | 2.50 | 2 | 2.01 | 2 |
| Business | average | 1.79 | 1 | 1.72 | 1 | 1.71 | 1 |
| | relatively large | 0.21 | 0 | 0.16 | 0 | 0.21 | 0 |
| | large | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 |

higher than the median fifty percent of the time. The upper bound of the load curve is indicated by the maximum value. Higher maximum values mean more coinciding electricity demand.

Dishwashers consume more electricity than tumble dryers, which in turn consume more than washing machines according to the average powers in Table 5.8. The mean power demand of washing machines and tumble dryers is slightly higher when the total electricity demand is higher in the case of day-consumers. The exception is large day-consumers, where the average power demand of washing machines and tumble dryers is almost the same as for relatively large day-consumers. Relatively small day-consumers' dishwashers' average power demand is lower compared to the other day-consumers, where the average power demand of the devices is about the same.

The appliance load curves of the average weekday, the average weekend day and the average day of the year for the average day-consumer are depicted in Figure 5.4. The electricity demand of all wet appliances is higher during a weekend than during a weekday. The influence of night tariff (see Section 2.2) is clearly visible in the load curve of the dishwasher. Another trend is the seasonal effect. The power demand of wet appliances is higher during winter compared to summer.

### Validation

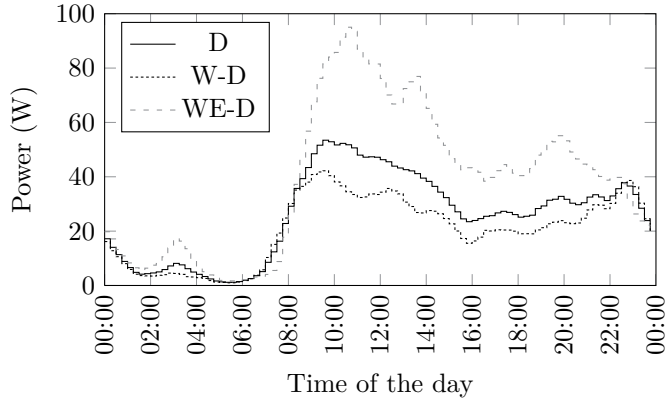
The load curves are validated by comparing the smoothed load curves of the average day consumer with these in [47]. The work of Stamminger et al. is based on surveys (Section 4.1.2). The average day-consumer is the largest group and the estimated total electricity demand is close to, but slightly lower, than the average total electricity demand of all customers. Because of the under-representation of small customers, the average day consumer is here used as representative for all customers.

Shape and the size of the washing machine load curve are comparable to the findings in [47], but shifted in time: the start in the morning is at 8 am compared to 4 am in [47]. The average power in [47] is higher, 37 W compared to 26 W (Table 5.8), a relative difference more than 40 %.

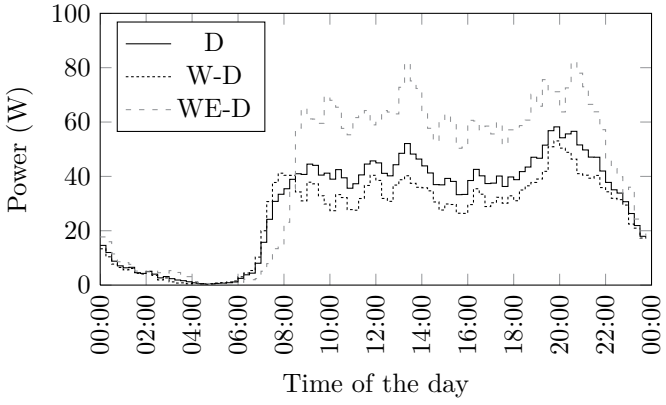
The average power demand of tumble dryers is around 103 W in [47], which is three and a half times higher than the average power demand of Table 5.8 (30 W). The shape of the load curves however, are almost identical. The difference in average power is mainly due to the ownership rate (34 % assumed, while 73.8 % [11] is more likely) (Section 4.1.2). 103 W on average means an

Table 5.8: Wet appliances: power demand according to Linear measurement data

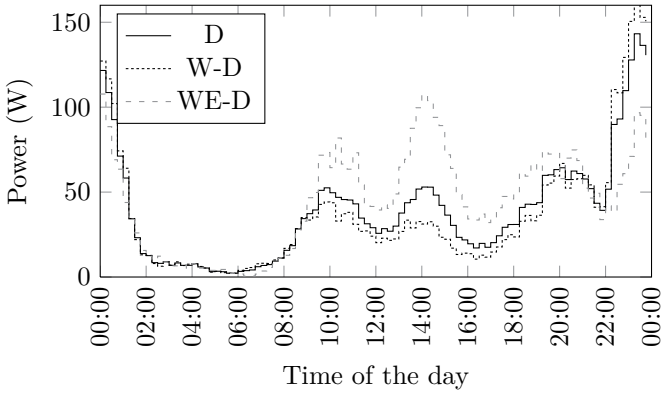
| Type | Sub-type | Washing machine | | | Tumble dryer | | | Dishwasher | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | mean | median | max | mean | median | max | mean | median | max |
| Day | relatively small | 21.75 | 16.17 | 106.48 | 25.18 | 23.86 | 103.90 | 30.90 | 25.34 | 135.61 |
| | average | 25.92 | 24.30 | 146.25 | 29.36 | 29.18 | 114.77 | 39.21 | 29.12 | 190.50 |
| | relatively large | 28.16 | 26.50 | 170.51 | 34.13 | 31.61 | 151.00 | 37.46 | 26.97 | 201.41 |
| | large | 28.56 | 25.79 | 200.70 | 33.99 | 29.57 | 156.36 | 37.16 | 24.42 | 248.49 |
| Night | large | 25.60 | 22.64 | 136.70 | 28.46 | 25.18 | 134.12 | 35.28 | 21.38 | 206.36 |
| Business | average | 27.68 | 24.45 | 214.23 | 33.09 | 27.96 | 169.61 | 41.52 | 19.36 | 410.16 |




(a) Washing machine




(b) Tumble dryer




(c) Dishwasher

Figure 5.4: Wet appliance load curves for the average day-consumer

electricity demand of around 900 kWh on a yearly basis, while around 250 kWh (29.4 W on average) seems more realistic.

The shape of the dishwasher load curve differs from [47]. The dishwasher is often started after every meal and mostly at the start of night tariff as shown in Figure 5.4c. However, in [47], dishwashers are mainly started on late afternoon and on evening. The average power demanded by a dishwasher is around 49 W in [47], while the dishwasher of the average day-consumer demands 39 W. The relative difference is around 25 %.

## 5.4 Conclusions

Electricity demand is described according to the total annual electricity demand and the demographic properties related. The distribution of the total annual electricity demand makes it possible to compare a customer to the whole population. Multiple distributions have been tested to fit the distribution and a Weibull function performed best, i.e. lowest $\chi^2$-error. The mean total annual electricity consumption is 4.9 MWh. Most people, i.e. the mode, have an electricity consumption of 2.8 MWh. The median (3.6 MWh) is used by the regulator to indicate the 'average' consumer.

The mapping of demographic properties to electricity demand is done by applying machine learning algorithms. The combination of demographic properties predicting electricity demand best, i.e. highest receiver operating curve (ROC) and lowest type I and type II errors, are regarded the most discriminating factors. Three machine learning algorithms are applied to map various combinations of properties onto demand. The algorithm with the best performance, again highest ROC and lowest errors, is used for the mapping. The decision tree approach (C4.5) performs better than the neural network (Multilayer Perceptron) and the Support Vector Machine (Sequential Minimal Optimisation) approach. The discriminating factors are found to be age, number of inhabitants, number of inhabitants who are at home during the evening, surface area for business, ownership of a freezer, ownership of a dishwasher and housing type.

Customers with a similar electricity demand are grouped to find the trends in demand. Expectation maximisation is a theoretically sound clustering algorithm that can be adapted to scale data up. The data up-scaling is required to spread customer data over the various found customer groups. The number of clusters has been limited to ten, double the number of day consumer types defined by the regulator and low enough to limit the number of outlier groups. Ten groups (the maximum number allowed) are found, with three subgroups: day, night

and business consumers, distinctions also made by the regulator. Each group is named according to timing and magnitude of their consumption and whether or not the group contains business consumers.

The clustering algorithm is adapted to obtain fuzzy cluster membership, is called the relaxed cluster membership. The advantage of the relaxed cluster membership is the possibility to scale up data. Households are part of multiple clusters instead of just one cluster, hence more households per cluster. The disadvantage is that the distributions change slightly and customer type probabilities differ. The relaxed cluster membership is corrected by including the original cluster probabilities. The corrected results produce distributions and cluster probabilities closer to the original ones.

The electricity demand of wet appliances is important for the estimation of the potential for active demand of those wet appliances. Load curves of the wet appliances' electricity demand per customer type are created from the measurements for the first estimates (dataset 1). The relaxed EM clustering algorithm determines the cluster membership based on the electricity demand at the connection point. The weights are used to scale the electricity demand of customers and to build the appliance electricity demand per customer type.

The resulting load curves of the average day-consumer have been compared to the literature. The shape of the load curves of washing machines and tumble dryers are similar to these described in the literature. The load curves of dishwashers on the other hand are influenced by the tariff structure in Belgium and hence differ from these in the literature (Germany and France combined).

Dishwashers consume more electricity than tumble dryers, which on their turn consume more than washing machines according to the average power demand of the wet appliances. The total demand for electricity is correlated with the average power of wet appliances. More inhabitants results in more overall electricity demand and also in a higher usage of the wet appliances, which explains the relation between total electricity demand and electricity demand by wet appliances.

# Chapter 6

# Privacy

An overview of a household situation can be derived from a limited number of appliances, mainly those with high power rating. The average number of washes per week is for example correlated with the number of household members. The time at which appliances are used gives insight on the life pattern of the household. To obtain a detailed view of the household situation, multiple detected appliances are required [138].

The detection of appliances in the electricity demand at the connection point is explained in Section 6.1. High resolution data makes it possible to detect appliances. With the resolution decreasing, the difficulty to detect appliances increases.

Certain households have submetering infrastructure, for individual appliances. Detailed appliance setting can be derived from submetering data with a resolution as low as 15 minutes, as presented in Section 6.2.

## 6.1 Appliance detection in household demand

Each appliance has its own load signature, depending on the components [53]. The detection of the components of an appliance enables the appliance detection. Variations on the load signature of appliances can be detected if the data resolution, i.e. frequency of measurements, is high enough.

### 6.1.1 High resolution data

The apparent power at the connection point of an apartment is presented in Figure 6.1[1]. The resolution of the data is 6 seconds, considered a relatively high resolution for non-intrusive appliance load monitoring (Section 4.2). Three high power appliances are detected in the 4 hour time frame: washing machine, instant coffee machine and oven.

The components to distinguish a washing machine from other appliances are the heating resistor and the motor. The power of a washing machine heating resistor ranges from 1.8 kW up to 2.5 kW typically being 2 kW. The power of a drum motor of a washing machine ranges from 300 to 700 W. In 2005, 94 % of the drum motors were universal motors [139]. The 'speed' of the universal motors are controlled using phase cutting, with a TRIAC (triode for alternating current), resulting in distortion in active and reactive power. Other possible motors are three phase induction motors (1 % in 2005) and brushless DC motors (5 % in 2005) [139]. The combination of 2 kW active power combined with distortion means the use of a washing machine. However, currently, the share of brushless DC motors is higher.

The washing machine cycle is indicated in Figure 6.1. The jump in electrical power indicated by number 1 is combined with distortion, which indicates the use of a washing machine. The second jump in electrical power (number 2) points again to the use of a heating resistor. The first jump in electrical power is most likely to be a prewash because of the combination of small electrical energy demand in the first jump and the length of the cycle. Thus the second jump in electrical power is part of the washing machine cycle. More information about the operating principle of a washing machine is given in Section 6.2.1.

Coffee machines have a heating resistor with a power of 0.6 up to 2 kW. The thermal inertia is low, which results in a high frequency of switching on and off [53]. A coffee machine remains off for a longer period than it stays on: heating only restarts when the water gets cold. Instant coffee makers have a different load signature. The water in the small reservoir gets preheated when the device is switched on. The next jump is when the programme gets executed, which is in general not long after the device is switched on. When the coffee is made, the machine starts heating the reservoir again (Figure 6.1).

Electrical ovens have a power demand of 1.8 to 6.5 kW, differing from stoves in the on/off time. An oven remains on and switches off in several minutes as time frame. Stoves on the other hand switch on and off in several seconds [53]. The jumps in power in Figure 6.1 could have been from a tumble dryer. However,

[1]The data is measured at the point of common coupling of the apartment of a friend, using a Flukso-meter.

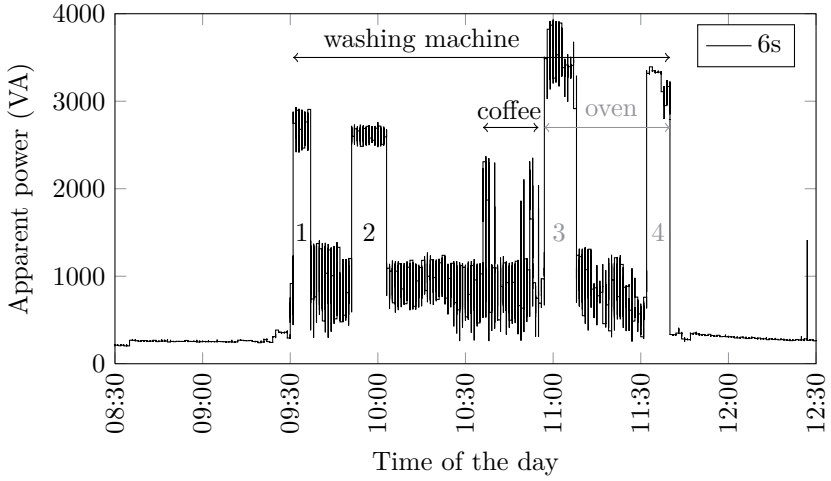



Figure 6.1: Apparent power at the connection point of an apartment

tumble dryers have a power demand at the end of the cycle to cool down the laundry, which is not observable in the load profile of Figure 6.1. Both jumps in power (numbers 3 and 4) are caused by an oven.

### 6.1.2 Lowering the resolution

The impact of lowering the data resolution is shown in Figure 6.2. The presented measurement intervals are 30 seconds, 5 minutes and 15 minutes. Appliances with a high on/off switching frequency become undetectable. For instance, the instant coffee machine is undetectable in the 5 minute data.

The washing machine is still visible in the 5 minute data, but the distortion is gone. Washing machines are hence hard to distinguish from tumble dryers for example. The same holds for the oven.

Figure 6.2 visualises what is explained in Section 4.2: the difficulty in detecting appliances rises when the resolution becomes lower. On a fifteen minute resolution, detecting appliances becomes very hard: maximum accuracy is 59 % in [60], other work reported the detection of air conditioners [57]. However no follow up studies were found.

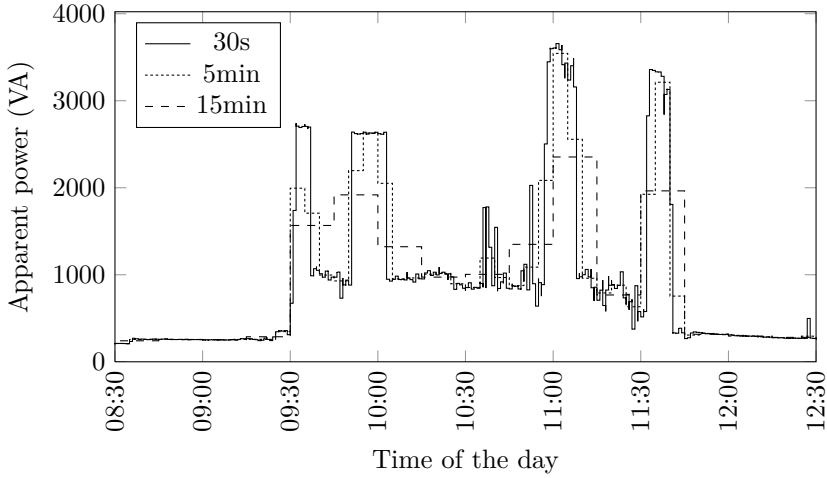



Figure 6.2: Measurement data with various resolutions

## 6.2 Appliance settings detection in sub-metering data

The detection of appliances in sub-metering data helps to model use and corresponding electricity demand of appliances. The operating principles have to be understood first, helping to create cycle and settings detection algorithms, enabling to estimate private information.

The appliances' metering data has a resolution of 15 minutes (Section 2.5). Cycle detection and settings estimation algorithms are applied to the project's dataset 1 (Section 2.5.1). The appliances' measurements and programmes' logging of the field test (Section 2.5.2) help to make privacy estimations.

### 6.2.1 Washing machine

The electrical load of a washing machine cycle depends on the programme, the load and the energy class. The programme itself is determined by the selected temperature, the (extra) water demand and the spinning speed. The operating principles, together with the cycle detection and the settings estimation algorithm for the basis to estimate privacy information such as 'expensive' clothing and household expansion.

The cycle detection algorithm detects the start and the end of a cycle in the load profile of a washing machine. The stand by power of washing machines is low (approximately 5 W). The cycle detection algorithm performs well if the sum of the electricity consumption of the cycles plus the consumption in stand by is equal to the total.

The found cycles are the input for the settings estimation algorithm. The settings to be estimated are programme, amount of water to be heated, laundry weight, and temperature. The described algorithms will be used to model washing machines in Section 7.3.2.

### Operating principle

A typical washing cycle starts by passing water through the detergent tray and letting the water into the drums of the washing machine. The water is heated to the required temperature. The inner drum of the washing machine rotates the laundry back and forward to remove all dirt. Once the laundry is clean, the water is let out of the machine and fresh water is taken in. The fresh water rinses the detergent, by rotating the inner drum back and forward. The water is let out of the machine before swinging the laundry dry [48].

The most energy intensive part of a cycle is heating the water to the required temperature. The water inlet temperature is typically 15 ℃. The power of the heating resistor ranges between 1.8 kW and 2.5 kW and is usually 2 kW. About 15 litres of water needs to be heated for 7 kg of laundry. The amount of water scales with the laundry weight. Some programmes, for example delicates, require double or triple the normal amount of water to be heated. The efficiency of heating is around 85 %, though the efficiency varies between energy classes. The temperature settings of a washing machine are 30 ℃, 40 ℃, 60 ℃ and 85 ℃.

The rated power of the drum motor ranges from 300 W to 700 W. In 2005, 94 % of the motors were universal motors, three phase motors were used in 5 % of the washing machines and brushless DC motors were found in the remaining 1 % [139]. However, the latter is gaining popularity.

Washing machines have a base load during the execution of a cycle, the base load varying between 100 W and 120 W, depending on the washing machine type and brand. The duration of a cycle scales with the amount of laundry: a 6 kg programme takes around 2 h, a 3 kg programme only requires 1 h. Half of the heating time needs to be added to the respectively 2 and 1 hour: a cycle with a 60 ℃ temperature setting takes longer than a cycle with a setting of 30 ℃.

### Cycle detection

A washing machine cycle has four different states: switched off, base load, heating and spinning. The states are based on the operating principle. Spinning is hard to distinguish from base load in fifteen minute data: the drum motor only spins for some minutes at full speed or spins at lower than rated power. Therefore, the spinning state is discarded in the cycle detection.

The threshold for the base load state is set at 30 W. The value is smaller than the 100 W of power demand during base load, because 'base load' is often demanded for less than 15 min. 30 W corresponds with 4 min of base load.

The heating state is found at the beginning of a washing machine cycle. The detection of a cycle is hence done by the heating state. For certain programmes, the duration of heating is short. The threshold for the heating state is therefore set at 400 W (300 W heating, which corresponds with 2 min 15 s of heating, plus 100 W base load). The value (400 W) is high enough to not confuse with spinning.

**Algorithm 6.2.1:** CycleDetect($profile$)

```
started ← false ; step ← 0; cycles ← new Array
for each state in profile
do { step ← step + 1
     if state is heating
       then { if started and not previous state is heating
                then { end ← step
                       cycles.append((start, end))
                       start ← step
              else if previous state is base load
                then {start ← (step − 1)
                else {start ← step
              started ← true
     if state is off and started
       then { end ← step
              EstimateSettings(start, end)
              started ← false
return (cycles)
```

Algorithm 6.2.1 describes how washing machine cycles are detected. The two most important control points are the 'heating' and the 'off' state. A 'heating'

state indicates a possible cycle start, while an 'off' state refers to a possible cycle ending.

If the washing machine is already started when encountering a new 'heating' state and the previous state was not 'heating', but 'base load', the previous cycle is parsed, a new cycle is started and the new start is set at the current step.

Otherwise, a cycle is started without having to parse a previous one. The start depends on the previous state: if the previous state was 'base load', the cycle is started at the previous state.

A cycle is ended and parsed when a new cycle is found or when an 'off' state is detected during the current cycle.

### Settings estimation

The energy required to heat the water to the required temperature depends on the amount of water and the required temperature. Therefore, an estimation of the amount of water to be heated and the temperature is needed to parametrise the electricity demand of washing machines.

The settings estimation algorithm works iteratively. An estimation of the laundry weight and the amount of water to be heated is made based on the total duration of the cycle. The weight and the corresponding water together with the heat demand is used to calculate the temperature offset. The closest temperature setting given the water inlet temperature and the temperature offset is selected. The amount of water to be heated and the corresponding weight of the laundry is updated according to the new temperature. Most of the settings detection is done experimentally.

The cycle detection algorithm passes the start and the end of a cycle to the settings estimation algorithm. The duration (in hours) of a cycle is the end-step minus the start-step divided by four (a step is 15 minutes long). The duration is slightly corrected for starts and ends that occur half way the fifteen minutes: if the power of the first or of the last step is less than 40 W, the duration is reduced by 0.1 hour (6 minutes).

A first estimate of the weight of the laundry ($m_l$) is done by multiplying the duration (in hours) by 3, given that a 3 kg programme generally requires 1 hour to finish. The amount of water to be heated scales with the weight of the laundry: 15 litres of water are needed for 7 kg of laundry. The weight multiplied by 15 l and divided by 7 kg results in the initial guess of the amount of water to be heated ($m_w$).

Some programmes require a higher amount of water: 'delicates' (likely to be the selected programme if 2 kg $\leq m_l <$ 3 kg), requires double the amount of water to be heated. Also if the calculated weight is lower than 1.5 kg, wool is a likely programme, hence the initial amount of water to be heated is tripled.

The electrical energy used to heat the water to the appropriate temperature ($E_h$) is determined by subtracting 'base load' electricity demand from the total electricity demand during heating steps. The effective heat ($Q_h$) is lower than the electrical energy ($E_h$) because of the efficiency ($\eta_{eff}$) of 85 %. The heating energy is the basis to estimate the temperature setting ($T_{est}$): the inlet temperature of the water of 15 ℃ is raised by the temperature increase of heating up the water ($\Delta T$). The temperature increase is determined by the specific heat coefficient ($c$) of water (4200 J/kgK $\equiv$ 7/6 Wh/kgK),

$$\Delta T = \frac{Q_h}{m_w \cdot c} = \frac{E_h \cdot \eta_{eff}}{m_w \cdot c} \tag{6.1}$$

The temperature setting ($T_{sel}$) closest to the estimated temperature ($T_{est}$) is chosen as the selected temperature. Based on the selected temperature, the water demand ($m_w$) and the laundry weight ($m_l$) are recalculated,

$$m_w = \frac{E_h \cdot \eta_{eff}}{(T_{sel} - 15) \cdot c} \tag{6.2}$$

Some corrections are needed. If the temperature is higher than 80 ℃, the amount of water is doubled because a cycle with more water demand and lower temperature is more likely. If the newly calculated temperature is higher than 60 ℃ and the weight is lower than 4 kg, the amount of water to be heated is set to be three times the initial water demand: double water demand and high temperatures are less likely again.

Next to the temperature check, a weight check is done. If the estimated weight of the laundry is higher than 7.7 kg, the temperature is raised (for example from 30 ℃ to 40 ℃); except, if the temperature is already 85 ℃, then the amount of water is increased.

To check whether the results are good, the recalculated is compared to the measured electricity demand. If the relative difference between both is more than 20 %, the amount of water to be heated is raised (from normal to double or from double to triple) and the settings are recalculated. The new settings are considered to be the correct ones when the newly calculated is closer to the measured electricity demand.

### Validation

Project's dataset 1 (Section 2.5.1) contains measurements of 88 washing machines with cycles detected in 82 of those load profiles. 10 of those profiles are randomly selected to validate the cycle detection and the settings estimation algorithm (Table 6.1).

The measured load profiles of the washing machines contain gaps. For only 63 washing machines, a load curve of the average week for each quarter of the year could be constructed (Section 2.5.1). Therefore, the percentage of time without measurements is included and indicated as 'no data'. The total measured electricity demand of the load profile is shown in the 'total' column. If the average power during in the fifteen minute step data is below 5 W, the washing machine is considered to be in 'stand by'. The total demand during 'stand by' is also expressed in the table.

To validate the cycle detection algorithm, the electricity demand during the detected cycles is calculated and compared to the total demand minus 'stand by'. The detected electricity demand is close to the difference between 'total' and 'stand by'. The threshold of the 'base load' state in the cycle detection algorithm is set to 30 W, which differs from the 5 W 'stand by'. For washing machine 32 for example, the electricity demand for power between 5 and 30 W is 1.80 kWh. However, this means that for washing machine 32, 7.47 kWh remains unallocated. For washing machine 67, there is even 9.07 kWh unallocated. At certain moments in time, average power demand during 15 minutes is higher than the 'base load', but lower than the 'heating' threshold.

The settings detection algorithm is validated by comparing the original electricity demand of the cycles with the calculated using the found settings. As Table 6.1 shows, the largest relative difference between the total of the detected and calculated cycles is in the case of washing machine 66, namely 3.28 %.

Not only the total demand is relevant, but also the difference per cycle. To compare those, the root mean square deviation (RMSD) and the coefficient variation of the root mean square deviation (CV(RMSD)) are calculated. The RMSD is calculated by taking the summation of from calculated values $\hat{u}$ subtracted from measured values $u$, divided by the number of values $n$ (6.3) and represents the standard deviation of the differences between calculated and measured values. CV(RMSD) is determined by dividing the RMSD value by the average of the sample $\overline{u}$ (6.4) and expresses the inverse of the signal to noise ratio SNR (6.5).

Table 6.1: Validation of washing machine cycle detection and settings estimation algorithm

| wm nr. # | no data [%] | cycles # | total [kWh] | stand by [kWh] | detected [kWh] | calculated [kWh] | RMSD | CV(RMSD) |
|---|---|---|---|---|---|---|---|---|
| 5 | 5.36 | 242 | 130.40 | 2.07 | 126.28 | 128.36 | 0.0317 | 0.0607 |
| 8 | 73.57 | 9 | 8.67 | 0.47 | 8.16 | 8.20 | 0.0382 | 0.0421 |
| 13 | 69.79 | 60 | 46.41 | 0.17 | 46.19 | 46.33 | 0.0396 | 0.0514 |
| 25 | 93.28 | 10 | 8.10 | 0.07 | 8.01 | 7.99 | 0.0345 | 0.0431 |
| 28 | 65.25 | 184 | 136.22 | 2.12 | 131.95 | 135.79 | 0.0471 | 0.0657 |
| 32 | 12.46 | 596 | 336.86 | 3.68 | 323.91 | 321.51 | 0.0423 | 0.0778 |
| 42 | 20.22 | 172 | 163.59 | 0.19 | 163.29 | 166.63 | 0.0380 | 0.0400 |
| 45 | 7.75 | 343 | 232.54 | 8.54 | 223.50 | 220.40 | 0.0302 | 0.0463 |
| 66 | 2.18 | 526 | 432.70 | 1.88 | 428.04 | 442.07 | 0.0452 | 0.0555 |
| 67 | 40.42 | 140 | 70.67 | 0.57 | 56.49 | 55.72 | 0.0283 | 0.0701 |

| Household size | 1 | 2 | 3 | 4 | 5 | $\geq 6$ |
|---|---|---|---|---|---|---|
| Washes per week | 2.1 | 3.4 | 4.9 | 6.4 | 6.3 | 7.0 |

Table 6.2: Relation between household size and average number of washes per week.

$$RMSD = \sqrt{\frac{\sum_{j=1}^{n}(\hat{u}_j - u_j)}{n}} \tag{6.3}$$

$$CV(RMSD) = \frac{RMSD}{\overline{u}} \tag{6.4}$$

$$SNR = \frac{1}{CV(RMSD)} \tag{6.5}$$

The absolute error (RMSD) is small (less than 50 Wh per cycle for the washing machines in Table 6.1) and comparable for the various customers. The relative error (CV(RMSD)) is small as well. Signal to noise ratio expresses the ratio of useful versus irrelevant data. A ratio value of more than one corresponds to more signal than noise, one of more than five is regarded to be good. Customer 42 has the highest (25.0) and customer 32 the lowest (12.9) signal to noise ratio.

### Privacy

The size of a household influences the number of washing machine cycles per week. Table 6.2 shows how the two are related [48]. An increase in the number of washes per week hence indicates an increase in the number of persons in the household. If the increase is combined with more activity during the night it is possible that the inhabitants became parents. As mentioned in Section 4.2, becoming a parent is one of the major interest points for advertisers.

An adaptation of the settings detection algorithm makes it possible to detect 'special' programme cycles such as 'delicates', 'woollens' and 'shirts'. The programmes are used for more expensive clothes. To show the possibilities of detecting privacy sensitive information, two rules are added to the algorithm to detect 'special' programmes. A cycle is detected as 'delicates' or 'shirts' if the weight ($m_l$) in the initial loop is in between 2 kg and 3 kg. A laundry

| TPR | FPR | Precision | Accuracy |
|---|---|---|---|
| 53.57 % | 10.20 % | 50.00 % | 84.00 % |

Table 6.3: Evaluation of special programme selection.

weight smaller than 1.5 kg indicates 'woollens'. In between the 'delicates' and the 'woollens' programme, the cycle is considered to be 'express'.

The 'special' programmes are marked in a set of randomly washing machine profiles. 104 washing machine load profiles are available in the project's dataset 2 (Section 2.5.2). 48 profiles have at least one special programme marked in the configuration logging and accompanying data. 10 profiles are randomly selected, resulting in 175 cycles with accompanying programme information.

Table 6.3 shows how the two simple rules perform in the detection of 'special' programmes. In total 28 of the 175 cycles are special programmes. The rules are able to find 15 of them, the 13 others are not marked. Also 15 other cycles are marked as special by the rules while they are not. The accuracy is of the algorithm is high because of the high number (147 in total) of 'non-special' programmes. The false positive rate FPR (5.6) is low. The true positive rate TPR (5.5) and the precision (5.7) are mediocre.

The results are influenced by one load profile with 25 cycles where the rules performed badly. Neglecting this profile, the accuracy would have been 90 %, the true positive rate 65.2 %, the false positive rate 5.5 % and the precision 68.2 %, being good to very good results. Cycles started directly after the previous one are harder to distinguish. The algorithm sometimes misses cycles with a 'heating' step with an average power during fifteen minutes which is smaller than 400 W. Also, the two rules do not capture all possibilities of 'special' programmes.

Extra rules should be added to improve the detection of special programmes. However, the aim was to show that privacy information could be obtained from measurements with a low resolution (15 minutes) and simple additional rules.

### 6.2.2 Dishwasher

The programme, the plate settings and the energy class are the determining factors of the electrical load of a dishwasher cycle. The programme can require extra water and has a temperature setting for washing and hot rinsing. Income and other privacy information can be estimated with the help of the detection

and the number of cycles of the appliance. The cycle detection and the settings estimation algorithm are used in Section 7.3.4.

### Operating principle

Dishwashers work in three steps: washing, cold and hot rinsing. The tub of the dishwashers gets filled with fresh water in each of the steps. The water gets heated to the required temperature during washing and hot rinsing. Rotating spray arms ensure that the water is spread around on the dishes.

A dishwasher with a 12 plate setting (A class) has a water demand of around 15 up to 22 litres [47, 48]. The water demand scales with the plate setting. A dishwasher with 6 plate setting only consumes half of the water of the 12 plate setting. Half of the heated water is used for washing, the rest for hot rinsing.

Water heating is, just as for washing machines, the most energy intensive part of a cycle. The power of the heating resistor varies between 1.8 and 2.5 kW, with 2 kW as most frequently used [47]. The heated water heats the dishes as well, which results in an efficiency of only 55 % for the heating of the water. The temperature during washing is often 5 to 10℃ lower than during hot rinsing. However, for simplicity of the model, this is not taken into account.

The washing step starts directly after the device starts, except if pre-rinsing is selected, taking about 10 minutes to finish. The heating resistor does not operate constantly: the resistor is switched on and off in pulses of several minutes. The average power demand during washing is around 1.5 kW, but varies between manufacturers.

Cold rinsing ensures the further cleaning of the dishes and takes, in general, a half hour. However, some programmes have up to 1 hour of cold rinsing. The electricity demand is the base load of approximately 100 W.

Hot rinsing is required to get the dishes dry. The hot water sprayed onto the dishes, heats the dishes. The temperature inertia of the dishes ensures that the dishes dry. The disadvantage of this approach is that plastic does not dry as well as glass, ceramic or metal. Some dishwashers blow therefore air through the heating resistor and the dishes for a short while. The latter is not included in the model.

### Cycle detection

Dishwasher cycles have four different states: ‘washing’, ‘cold rinsing’, ‘hot rinsing’ and ‘off’. The states are translated into the detectable states ‘heating’,

'base load' and 'off'. Heating covers 'washing' and 'hot rinsing', while 'base load' is mainly intended for 'cold rinsing'.

The thresholds for the different states are the same as the ones used for the washing machine. If the power demand during 15 minutes is higher than 30 W, the dishwasher is considered to be in 'base load'. The threshold for heating is set at 400 W.

Algorithm 6.2.2 represents the dishwasher cycle detection algorithm. 'Off' and 'heating' states are again the most important ones for detection purposes. 'Heating' indicates a possible cycle start, while 'off' means a possible cycle end.

If the dishwasher has not been started yet and a 'heating'-state is encountered, the dishwasher is started. The exact start depends on the previous state: if it is 'base load', the start occurs during the previous state, otherwise the start is at current state.

The algorithm could miss a new cycle start which occurs directly after another state. However in practice, people do not start their dishwasher less than 30 minutes after the end of the previous cycle [48].

**Algorithm 6.2.2:** CycleDetect($profile$)

$started \leftarrow$ **false** ; $step \leftarrow 0$; $cycles \leftarrow$ new Array

**for each** *state* in $profile$
**do**
- $step \leftarrow step + 1$
- **if** *started*
  - **then**
    - **if** *state* is off
      - **then**
        - $end \leftarrow step$
        - $cycles.append((start, end))$
        - $started \leftarrow$ **false**
  - **else**
    - **if** *state* is heating
      - **then**
        - $started \leftarrow$ **true**
        - **if** previous *state* is base load
          - **then** $\{start \leftarrow step - 1$
          - **else** $\{start \leftarrow step$

**return** (*cycles*)

### Settings estimation

Heating water is energy intensive and requires most energy during a dishwasher cycle. The amount of energy depends on the amount of water and the

temperature setting. The water demand varies with the plate setting.

Dishwashers with a plate setting of 12 are most frequently sold, their market share is 83.3 %. The second and the third most frequently sold have 9 and 8 plate settings, with a respective market share of 8.5 % and 4.1 % [140]. Thus, 95.9 % of all dishwashers sold, have a plate setting of 12, 9 or 8. Therefore, only those plate settings are considered.

The amount of water that needs to be heated (both washing and hot rinsing) is assumed to be 13 litres for a 12 plate setting. The water demand scales with the plate setting: a 9 plate setting requires 9.75 l to be heated and a 8 plate setting 8.7 l.

The temperature settings differ between brands, some have 35 ℃ as lowest temperature, others 45 ℃. To compensate for the differences, four temperature settings are assumed: 40 ℃, 50 ℃, 60 ℃ and 70 ℃, both for washing and hot rinsing. The temperature values are based on the settings distribution in [47]: 40 ℃is the average of the 35/45 ℃ range, 50 ℃is the lower boundary of the 50/55 ℃range, 70 ℃remains 70. To make the values equidistant, steps of 10 ℃are used: 65 ℃ is used in stead of 60 ℃, the difference between 65 ℃ and 70 ℃ is not large enough to distinguish amongst them.

Initially, the plate setting is assumed to be 12 because of the large market share. If the total recalculated electricity consumption is too high compared to the detected electricity demand, the algorithm is executed again with a lower value. A relative difference of 20 % is translated in a plate setting of 9, a relative difference of 40 % results in 8.

The electrical energy needed to heat up the water is calculated similarly as for the washing machine. However, the efficiency ($\eta_{eff}$) is only 55 % for dishwashers.

### Validation

The cycle detection is validated by comparing the sum of the electricity consumption of the cycles and the consumption during stand by with the total consumption, both should be equal. The validation of the settings estimation algorithm is done by comparing the electricity demand of the individual cycles with the calculated demand of the settings estimated. The sum of the detected cycles has to equal the sum of the recalculated cycles. Also the absolute (RMSD) and the relative difference (CV(RMSD)) between the individual cycles has to be compared and should be as low as possible.

The validation is done on the project's dataset 1 (Section 2.5.1). The data contains measurements of 53 dishwashers. In 50 of those, dishwasher cycles are

detected. 10 load profiles are randomly selected for validation purposes.

The 'no data' column shows the unavailability of the data during the year. The load profile of dishwasher 23 contains large gaps. However, there are 226 cycles detected. The profile is checked manually and the customer pulled the plug of the dishwasher some minutes after each cycle. The total electricity demand of each load profile can be found in the 'total' column. Average fifteen minute power below 5 W is marked as 'stand by' power.

Validation of the cycle detection algorithm is done by comparing the 'detected' electricity demand to the 'total' minus the 'stand by' electricity demand. Both are not completely the same but very close, as shown in Table 6.4. The explanation is the same as for washing machines: 'base load' is detected at 30 W while 'stand by' stops at 5 W.

The recalculation of the cycles allows for a validation of the settings estimation algorithm. The total electricity demand of the recalculated cycles for the different dishwashers are shown in Table 6.4. The settings detection algorithm overestimates the electricity demand in cycles. The largest relative difference between detected and calculated is 11 % (dishwasher 27), the smallest is 0.3 % (dishwasher 40).

The root mean square deviation (RMSD) is higher for dishwashers compared to washing machines and goes up to 214 Wh per cycle (dishwasher 25). The higher values for RMSD are partly explained by the higher electricity consumption per cycle: the relative error (CV(RMSD)) is higher for dishwashers as well, but the relative difference is smaller. To make the relative difference easier to compare signal to noise ratio (SNR, the inverse of CV(RMSD)) is compared as well. The highest SNR is found for dishwasher 9 (17.0), the lowest for dishwasher 25 (4.5). A SNR of 4.5 is only mediocre, the other SNR values are higher than 5 and considered to be good.

## Privacy

The detection of a dishwasher and its number of weekly cycles help to estimate the number of inhabitants and the total net income of the household.

The relation between household size and dishwasher ownership rate is shown in Table 6.5 [48]. The more inhabitants, the more likely the household owns a dishwasher. The frequency of using the dishwasher during the week is also correlated to the household size. A single person home has on average 2.6 dishwasher cycles per week, while a four person household operates the dishwasher on average 6.2 times [48].

Table 6.4: Validation of dishwasher cycle detection and settings estimation algorithm

| dw nr. # | no data [%] | cycles # | total [kWh] | stand by [kWh] | detected [kWh] | calculated [kWh] | RMSD | CV(RMSD) |
|---|---|---|---|---|---|---|---|---|
| 4 | 18.32 | 113 | 136.45 | 0.69 | 135.66 | 137.87 | 0.1018 | 0.0848 |
| 9 | 21.02 | 330 | 423.04 | 2.55 | 418.8 | 416.10 | 0.0747 | 0.0588 |
| 12 | 1.31 | 238 | 274.21 | 1.65 | 272.36 | 279.27 | 0.0815 | 0.0712 |
| 19 | 0.62 | 283 | 383.56 | 3.81 | 378.9 | 376.90 | 0.0898 | 0.0671 |
| 23 | 92.13 | 226 | 303.98 | 0.14 | 303.67 | 309.89 | 0.1034 | 0.0769 |
| 24 | 0.07 | 170 | 194.35 | 2.63 | 191.40 | 187.96 | 0.0789 | 0.0701 |
| 25 | 2.34 | 414 | 399.17 | 0.24 | 395.62 | 439.92 | 0.2140 | 0.2240 |
| 27 | 9.89 | 114 | 143.48 | 1.26 | 142.00 | 147.57 | 0.1141 | 0.0916 |
| 31 | 0.02 | 304 | 389.67 | 6.68 | 382.49 | 374.24 | 0.1039 | 0.0826 |
| 40 | 0.01 | 120 | 151.83 | 2.60 | 146.60 | 147.03 | 0.1042 | 0.0853 |

| Household size | 1 | 2 | 3 | 4 | $\geq 5$ |
|---|---|---|---|---|---|
| Ownership rate | 50 % | 66 % | 72 % | 78 % | 80 % |

Table 6.5: Relation between household size and dishwasher ownership.

| net income [€] | owners [#] | total [#] | ownership rate [%] |
|---|---|---|---|
| 0 - 1000 | 7 | 29 | 24 |
| 1001 - 2000 | 72 | 150 | 48 |
| 2001 - 3000 | 61 | 93 | 66 |
| 3001 - 4000 | 52 | 71 | 72 |
| 4001 - 5000 | 24 | 26 | 92 |
| over 5000 | 11 | 12 | 92 |

Table 6.6: Relation between net household income and dishwasher ownership.

61.1 % of the people participating in the survey (Section 2.4) have a dishwasher. The survey also polled for the net income of the household. 381 of the 491 households (i.e. 77.6 %) answered this question. 227 of those 381 households (59.6 %) have a dishwasher. The relation between the net income and dishwasher ownership is shown in Table 6.6. The higher the income, the more likely the family owns a dishwasher. A more formal way to describe the probability of the income given that he or she owns a dishwasher $P(income|dishwasher)$ is,

$$P(income|dishwasher) = \frac{P(income) \cdot P(dishwasher|income)}{P(dishwasher)} \tag{6.6}$$

where the probability of an income $P(income)$ is obtained from the Directorate General for Statistics and Economic Information [12], the probability of owning a dishwasher given the income is presented in Table 6.5 and the probability of owning a dishwasher $P(dishwasher)$ is also known, see above.

A single person home is more likely to have a lower net income than a household with two people. The household size and the net income have some relation. The combination of the information delivered by owning a dishwasher or not, the frequency of operation and information from other appliances gives insights of the household situation.

### 6.2.3 Tumble dryer

The load placed into a tumble dryer, together with the residual moisture level determine a tumble dryer's electrical load cycle. A small adaptation of the cycle detection and settings estimation algorithms enable detecting tumble dryers with a heat pump. Both help to relate the use of a tumble dryer to household income. The cycle detection and the settings estimation algorithms are the basis for the tumble dryer model in Section 7.3.3.

#### Operating principle

Tumble dryers evaporate the residual moisture in laundry. A fan blows air over the heating resistor and into the tub. The drum motor ensures that the laundry is swung back and forward so that the air isn't blown on the same pieces of clothing. After removing the moisture, the tumble dryer blows cold air over the laundry to cool it down. Modern tumble dryers use a heat pump instead of heating resistors.

The power rating of a heating resistor is typically 2 kW. Sometimes heating resistors of 3 kW or the combination of a 2 kW and a 1 kW resistor which can be operated individually are encountered as well.

The heating resistor is switched on the moment a cycle is started and operates until the laundry is dry. Laundry is considered to be dry when the humidity is at the desired level. Old tumble dryers use a timer.

#### Cycle detection

The states of a tumble dryer cycle are 'heating', 'base load' and 'off'. The tumble dryer heats when the heating resistor is switched on and 'base load' is demanded when cold air is blown. The states are the same as those of a washing machine cycle. Therefore, the same cycle detection algorithm as the one for a washing machine (Algorithm 6.2.1) is used. The thresholds for the different states are also the same.

#### Settings estimation

Tumble dryers evaporate the latent moisture in the laundry. The heat to remove the moisture ($Q_h$) depends on the amount of moisture and the initial temperature of the water. The electrical energy is related to the heat ($Q_h$) and the efficiency ($\eta_{eff}$), which is taken to be 90 %.

To determine the required heat ($Q_h$), the inverse reasoning holds. The electrical energy for heating ($E_h$) is the total electrical energy demand of the cycle ($E_t$), minus the electricity demand due to base load ($E_b$).

$$Q_h = E_h \cdot \eta_{eff} = (E_t - E_b) \cdot \eta_{eff} \tag{6.7}$$

The electricity demand because of 'base load' ($E_b$) is the 'base load' power (200 W) times the duration of the cycle $t_{cycle}$. The cooling of the laundry takes around 10 minutes.

$$E_b = P_b \cdot t_{cycle} \tag{6.8}$$

The maximum power encountered during the cycle and the total electricity demand are used to find the cycle duration ($t_{cycle}$) expressed in hours: the total electricity demand divided by the maximum power. Four powers are considered as maximum: 2 kW if the maximum power is higher than 1.8 kW, 1.5 kW when the maximum power is between 1.4 and 1.8 kW, 1 kW if the power is between 1.4 kW and 800 W and 500 W for powers between 400 and 800 W.

The amount of water in the laundry, i.e. the residual moisture, is calculated from the heat ($Q_h$). First, the water needs to be heated to 100 ℃. The water in the laundry is assumed to be at 15 ℃, resulting in a temperature increase ($\Delta T$) of 85 ℃. The specific heat of water ($c$) is 4200 J/kgK or 1.167 Wh/K. Thereafter, the water needs to be evaporated. The energy required to evaporate water is determined by the latent heat of water ($L$) being 2.26 MJ/kg or 6.278 kWh/kg.

$$m_w = \frac{Q_h}{c \cdot \Delta T + L} \tag{6.9}$$

Estimates about the weight of the laundry placed in the machine are done by making assumptions about the residual moisture level. The spinning speed of the washing machine cycle determines the residual moisture level. By combining the spinning speeds of France and Germany, an estimate about the spinning speed of Belgium is made. Due to a lack of data for Belgium, data of the neighbouring countries are used. 41 % of the washing machine cycles have a spinning speed between 1000 and 1300 rpm [48]. 1200 rpm and the accompanying 55 % residual moisture level are used to estimate the weight of the laundry ($m_l$).

$$m_l = \frac{m_w}{0.55} \tag{6.10}$$

### Validation

Measurements of 69 tumble dryers are available in the project's dataset 1 (Section 2.5.1). Tumble dryer cycles are found in 63 of them. 10 load profiles where tumble dryer cycles are detected, are randomly selected to validate the cycle detection and the settings estimation algorithms (Table 6.7).

The availability of the data per load profile is made clear by the 'no data' column. A high number of measurements are available compared to the randomly selected profiles of the washing machines and the dishwashers. The total measured electricity demand and the stand by power are again shown as respectively 'total' and 'stand by'. The upper boundary for stand by power remains 5 W.

The cycle detection algorithm is validated by comparing the cycles' detected electricity demand against the total demand minus the stand by electricity demand. The detection algorithm misses some of the cycles. For example tumble dryer 25 has 7.22 kWh which cannot be explained by stand by power. However, the cycle detection algorithm is able to capture most of the cycles: the largest relative difference between total minus stand by and detected electricity demand is 3.38 % (tumble dryer 28).

The recalculation of the electricity demand of the cycles is the basis to validate the settings estimation algorithm. The largest relative difference between detected and recalculated cycles is low (1.42 % for tumble dryer 19). The total recalculated electricity demand is higher than the detected demand.

The root mean square deviation (RMSD) of the detected versus the recalculated cycles is small. The highest value is encountered with tumble dryer 25: 27 Wh per cycle. The relative error (CV(RMSD)) is low, which results in high signal-to-noise ratios SNR of 60 to 250, knowing that the threshold for SNR is 5.

### Privacy

The ownership rate of a tumble dryer is related to the household size and the total net income of the household, both are visualised in respectively Table 6.8 and Table 6.9. The source for both tables are data from the survey of Section 2.4. Owning a tumble dryer hence gives information on household size and the total net income. The income estimation $P(income|dryer)$ is done in the same way as for dishwashers (6.6): the ownership of dishwashers is replaced by the ownership of tumble dryers.

However, to get a clear overview of the household size, the information needs to be combined with information from other appliances and total electricity

Table 6.7: Validation of tumble dryer cycle detection and settings estimation algorithm

| dw nr. # | no data [%] | cycles # | total [kWh] | stand by [kWh] | detected [kWh] | calculated [kWh] | RMSD | CV(RMSD) |
|---|---|---|---|---|---|---|---|---|
| 3 | 18.09 | 306 | 532.49 | 0.11 | 529.52 | 531.59 | 0.0154 | 0.0089 |
| 4 | 6.64 | 121 | 143.04 | 0.10 | 142.23 | 143.42 | 0.0168 | 0.0143 |
| 19 | 0.51 | 172 | 273.00 | 0.37 | 271.89 | 275.81 | 0.0262 | 0.0166 |
| 22 | 3.15 | 262 | 457.63 | 1.44 | 454.97 | 456.83 | 0.0150 | 0.0087 |
| 25 | 13.68 | 254 | 447.58 | 1.79 | 438.67 | 443.36 | 0.0274 | 0.0159 |
| 28 | 16.11 | 71 | 178.23 | 0.03 | 172.37 | 173.12 | 0.0157 | 0.0067 |
| 35 | 24.97 | 15 | 48.34 | 0.02 | 48.31 | 48.37 | 0.0128 | 0.0040 |
| 38 | 8.47 | 491 | 700.11 | 3.91 | 690.98 | 695.09 | 0.0164 | 0.0116 |
| 43 | 0.02 | 460 | 434.18 | 2.70 | 428.41 | 430.04 | 0.0135 | 0.0145 |
| 45 | 1.19 | 194 | 333.19 | 0.87 | 331.64 | 332.97 | 0.0156 | 0.0091 |

| Household size | 1 | 2 | 3 | 4 | $\geq 5$ |
|---|---|---|---|---|---|
| Ownership rate | 52 % | 70 % | 76 % | 88 % | 94 % |

Table 6.8: Relation between household size and tumble dryer ownership.

| net income [€] | owners [#] | total [#] | ownership rate [%] |
|---|---|---|---|
| 0 - 1000 | 13 | 29 | 45 |
| 1001 - 2000 | 99 | 150 | 66 |
| 2001 - 3000 | 70 | 93 | 75 |
| 3001 - 4000 | 55 | 71 | 77 |
| 4001 - 5000 | 23 | 26 | 88 |
| over 5000 | 10 | 12 | 83 |

Table 6.9: Relation between net household income and tumble dryer ownership.

demand. The average number of cycles per week on the other hand only gives marginal information about the household, a single person household uses the tumble dryer on average 2.4 times per week, while a five person household only operates the tumble dryer once per week more.

Tumble dryers with a heat pump are more expensive. The dryers have a lower power demand, around 1 kW, compared to 2 kW for a regular tumble dryer. All households in the field test, see Section 2.5.2, are equipped with a tumble dryer with a heat pump.

A rule is added to the settings detection algorithm to find the type of tumble dryer. Tumble dryers operate with a heat pump if the maximum power during a cycle is lower than 1 kW for 80 % of the cycles. 89 of the 95 tumble dryers are correctly labelled as heat pump tumble dryers by this rule, an accuracy of 93.7 %. A simple rule makes it thus possible to find out if the tumble dryer is an expensive one.

## 6.3 Conclusions

The detection of appliances in the total electricity demand of the household is possible if the resolution is high enough, for example a 6 second interval. The detection of appliances becomes harder when the frequency of measurements goes down. On a fifteen minute scale, it becomes very hard to detect appliances. A fifteen minute scale is the resolution used in smart meters.

Providers of home management systems gather, next to measurements of the electricity demand at the connection point, measurements of individual appliances. The appliance cycles can easily be detected and the settings of the appliance during the cycles can be estimated. Cycle detection and settings estimation algorithms for washing machines, tumble dryers and dishwashers are described. The algorithms are used further on to model the appliances.

Privacy sensitive information can be estimated from the appliance ownership and usage. Special washes indicate more expensive clothing. Household sizes are related to the frequency of operating washing machines and dishwashers. The estimation of household size can be enhanced by combining estimation based on total electricity demand, appliance ownership and appliance usage. The net income of a household and the ownership rate of dishwashers and tumble dryers are related as well. Tumble dryers with a heat pump are more expensive than tumble dryers that operate with a heating resistor. Whether the tumble dryer works with a heat pump is easily detected with a simple rule.

# Chapter 7

# Electricity demand for simulations

Residential electricity demand data is well protected in Europe because of privacy concerns [141, 138] (Chapter 6). The number of residential electricity monitoring companies are limited and the companies are not keen to share data and usually provide aggregated electricity demand. Information such as peaks in individual load profiles, together with load factors are lost because of demand aggregation. The selection and the generation of load profiles is described in this chapter.

A set of selected customers itself is considered representative if it represents the average population of a geographic area, such as Belgium. In some simulations, demographic properties and attitude towards active demand are important to draw conclusions, therefore only customers who responded to the survey (Section 2.4) are considered to sub-sample from. Because of the non-response, the set is biased. A quota optimisation technique is proposed.

Electricity demand at the connection point of houses is modelled with the use of Markov models. Distribution fits are the basis for models describing electricity demand of wet appliances. The models are able to generate load profiles of respectively the connection point and wet appliances.

## 7.1 Profile selection

Smart grid projects need a representative set of customers to draw conclusions for the whole population. However, people motivated and interested in smart grids are over-represented. The same problem occurs for the selection of profiles for simulations. If the sampling frame, i.e. the population to sample from, is not representative, it is difficult to sample a representative set of profiles with simple random sampling or systematic sampling. A method to cope with the selection problem is needed.

Various sampling techniques are compared in Section 7.1.1. An optimisation algorithm is used to select a set of households (Selection 7.1.3) according to information of the whole population (Section 7.1.2). The result presented in Section 7.1.4 is a set which fulfils the selection requirements.

### 7.1.1 Sampling techniques

The literature describes two classes of sampling techniques: probability and non-probability sampling (Section 3.4).

#### Probability sampling

The probability sampling techniques are simple random, systematic, stratified and cluster sampling. Simple random sampling and systematic sampling are not suited to fix the non-representativeness. Cluster sampling requires clusters of individuals to be selected. A possible clustering scheme is grouping households from the same village and selecting the village. However, this also doesn't fix the non-representativeness.

Stratified sampling makes it possible to obtain a sample representative of the target population. The sampling frame needs to be divided in distinct strata based on the relevant demographic properties and the total yearly electricity consumption. The probability of each of the strata has to be calculated. Individuals can be randomly sampled according to the probability of each stratum. A stratum is defined by the combination of different demographic properties and total electricity demand.

However, the Directorate General for Statistics and Economic Information does not provide information about the combination of demographic properties. It is for example unknown how many of the households are of size 3 in a semi-detached house and a surface area for business. The assumption that

demographic parameters are equally distributed in every combination could be made. For example, 30 % of the households are single person. Considering an equal distribution, this would mean that 30 % of the households with a low consumption as well as 30 % of the households with a high consumption are single person households. However, the assumptions would be wrong.

#### Non-probability sampling

The non-probability sampling techniques (Section 3.4) are convenience, modal instance, snowball, expert and quota sampling. Convenience sampling does not fix the non-representativeness. In modal instance sampling, only the mode is selected. Heterogeneity sampling samples diversity rather than in proportion to the target population. Snowball sampling requires individuals to suggest other individuals, which will result in similar demographic properties for the individuals.

Expert sampling requires an expert in the field to assemble the sample determining the demographic parameters for sampling. Machine learning processes (Section 5.1.2) are regarded as experts. Age of the respondent, number of inhabitants, number of inhabitants at home during the evening, surface area for a business, owning a freezer, owning a dishwasher and the housing type are hence regarded as the important demographic parameters.

Quota sampling resembles stratified sampling, but without the requirement for individual strata. The selection of individuals has to be done in proportion to the quota. The demographic parameters for which quota need to be defined are the ones found by the expert. The quota themselves are determined by information from the Directorate General for Statistics and Economic Information.

### 7.1.2 Quota definitions

The quota are related to the electricity demand itself and to the demographic parameters influencing demand. Total yearly electricity consumption is used as a proxy for the electricity demand during the year. The demographic parameters with quota are number of inhabitants per household and the housing type. Another requirement is that 63 % of the selected households need to have a connection for gas. The latter is a requirement from the Linear project.

#### Electricity demand

The total yearly electricity demand is described by a Weibull function in Section 5.1.1. The parameters of the fit are the same parameters for the cumulative Weibull distribution function. A cumulative distribution describes the probability that a variable $X$ is smaller than or equal to a value $x$.

The cumulative Weibull distribution is used to divide the distribution into eight equally probable classes. However, because of the use of bins, some are slightly more probable than others. The lower bounds of the classes are shown in Table 7.2. The upper bound of a class is the lower bound of the following. The lower bound of class 1 is 0 kWh. Class 8 does not have an upper bound.

#### Demographic parameters

The demographic parameters that predict total yearly electricity demand, electricity load factor and variance on power demand well are: age of the respondent, number of inhabitants and surface area for business [142]. Housing type is relevant for the total yearly electricity demand and variance (not for the load factor) and is hence considered as well.

The Directorate General for Statistics and Economic Information describes distributions over the demographic properties in Belgium and is considered to be an expert. The distributions of the number of inhabitants and the housing types are shown in Table 7.3 and 7.4 [12]. The social segmentation (attitudes) (Section 4.4.3) is also incorporated, however each segment got an equal probability.

### 7.1.3 Quota optimisation

The selection of households in proportion to the various quota is done with an optimisation algorithm. Constraint programming is used in both approaches to optimise the quota. The first approach uses logic programming, the second combines constraints with mathematical programming.

Constraint programming itself works in two levels. First, constraints are stated over the programming variables. Second, a computer program is written indicating how the variables should be modified in order to find the values of the variables satisfying the constraints [143].

Data representation

```
consumption(total, 1, 5410).
consumption(total, 3, 4567).
...
inhabitants(1, 3).
inhabitants(3, 2).
...
```

Rules and constraint definitions

```
get_ids_consumption(R, Min, Max) :-
  C $>= Min, C $< Max,
  findall(X, consumption(total, X, C), R).
...
housing_constraints(H) :-
  get_ids_housing(R1, 1), length(H1, 40), H1 :: R1,
  get_ids_housing(R2, 2), length(H2, 25), H2 :: R2,
  get_ids_housing(R3, 3), length(H3, 30), H3 :: R3,
  get_ids_housing(R4, 4), length(H4, 5), H4 :: R4,
  flatten([H1, H2, H3, H4], H),
  alldifferent(H).
...
```

Knowledge query example

```
housing_constraints(R1), consumption_constraints(R2), labeling(R1), permutation(R1, R2).
```

Figure 7.1: Constraint logic programming flow of quota optimisation

### Constraint programming

Constraint logic programming is a declarative programming approach combining logic and constraint programming. Logic programming is based on first order logic. Knowledge and rules are declared in the logic program. The logic program tries to answer a query based on its facts and rules. In constraint logic programming, queries are answered based on the stated knowledge, rules and constraints. The latter speed up queries [144].

The structure of the constraint logic program is shown in Figure 7.1. The data representation states the facts on the data, household 1 for example has an electricity consumption of 5410 kWh on yearly basis. Thereafter, rules and constraints are defined. `get_ids_consumption` couples the household-ids with an electricity demand between `Min` and `Max` to the variable `R`. `housing_constraints` requires that the distribution of the different housing types is met. The knowledge query starts the search for a solution.

Branch and bound algorithms are used when a query is requested. The algorithms may choose any household-id within the defined constraints. The constraint programming implementation has one large downside: the use of permutations. The elements cannot be ordered and thus permutations are

| distr. 1 | | distr. 2 | | | |
|---|---|---|---|---|---|
| | 1 | 1 | 2 | 3 | 4 |
| | 2 | 5 | 6 | 7 | 8 |
| | 3 | 9 | 10 | 11 | 12 |
| | 4 | 13 | 14 | 15 | 16 |

Table 7.1: Mathematical programming: bin numbering

required. Checking permutations increases the calculation time dramatically. An attempt to place the elements in an ordered form has been made by using a suspended quick-sort algorithm, but the constraint programming system orders the domains of the list elements instead of delaying ordering of complete list [142].

### Mathematical programming

Mathematical programming has been extended with constraint programming as well [143, 145]. A constraint can be converted into a mathematical function $f(x_1, x_2, \ldots, x_n)$ which results into 1 if the constraints $c(x_1, x_2, \ldots, x_n)$ are met.

The selection problem is translated into a mixed integer linear programming problem. Every distribution, i.e. every quotum is a dimension, in the vector space (Table 7.1). Combinations of bins are numbered and called combined bins. Combined bin 10 in Table 7.1 is the combination of bin 3 of the first and bin 2 of the second distribution. The number of combined $n_{cbins}$ bins is hence,

$$n_{cbins} = \prod_{d=1}^{n_{dists}} n_{bins}^{d} \tag{7.1}$$

The first constraint is the number of households per combined bin. The number of selected households has to be a natural number (an integer) and the maximum number of households to select from the combined bin $sel_{cb}$ is the number of households in the combined bin $n_{el}^{cb}$,

$$0 \leq sel_{cb} \leq n_{el}^{cb} \quad , \quad \forall sel_{cb} \in \mathbb{N} \quad \forall cb \in cbins \tag{7.2}$$

Table 7.2: Distribution and boundaries quotum classes total yearly consumption

| quotum class | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| lower bound [kWh] | 0 | 1277 | 2043 | 2809 | 3575 | 4342 | 5363 | 6896 |
| % required | 11.0 | 11.6 | 13.1 | 13.1 | 12.0 | 13.4 | 13.6 | 12.1 |
| # selected | 11 | 12 | 13 | 13 | 12 | 13 | 14 | 12 |

The total number of households to be selected $n_{el}$ is the second constraint,

$$\sum_{cb=1}^{n_{cbins}} sel_{cb} = n_{el} \tag{7.3}$$

The optimal solution is that each bin quotum per distribution $q_{d,b}$ is met for all distributions. The number of selected individuals per bin per distribution is quotum $q_{d,b}$ times number of households to be selected $n_{el}$. The real number of selected households per bin per distribution is however the sum over all combined bins related to the respective bin of the distribution,

$$\min_{x} \left( \sum_{d=1}^{n_{dists}} \sum_{b=1}^{n^{d}_{bins}} \left| q_{d,b} \cdot n_{tosel} - \sum_{cb}^{cbins_{d,b}} sel_{cb} \right| \right) \tag{7.4}$$

The result of the optimisation is the number of households to be selected per combined bin $sel_{cb}$, the selection within combined bins is done by uniform random sampling.

### 7.1.4 Selected profiles

The mathematical programming approach is executed for the selection of 100 households. The distribution of the number of households with respect to the total electricity consumption, together with the required percentages are shown in Table 7.2. The algorithm found a number per class corresponding to the rounded version of the required number. However, numbers might differ when an insufficient number of customers is available in a bin. The quota for number of inhabitants (Table 7.3) and housing types (Table 7.4) are met as well.

Table 7.3: Distribution of number of inhabitants per household

| inhabitants | 1 | 2 | 3 | 4 | 5 | $\geq 6$ |
|---|---|---|---|---|---|---|
| % required | 30 | 34 | 16 | 14 | 5 | 1 |
| # selected | 30 | 34 | 16 | 14 | 5 | 1 |

Table 7.4: Distribution of housing types in selection

| housing type | detached | semi-detached | terraced | apartment |
|---|---|---|---|---|
| % required | 40 | 25 | 30 | 5 |
| # selected | 40 | 25 | 30 | 5 |

## 7.2 Profile generation

The selection of load profiles is not always possible. Privacy issues or non disclosure agreements might prevent the use of original load profiles. A way to work with load profiles without needing the original ones is by modelling profiles. Modelling and generation of load profiles of the electricity demand at the connection of households is described in this section. The approach works top-down, in contrast to the bottom-up approaches described in the literature (Section 4.3.2).

The residential load data of Flanders as (Section 2.3) are the basis for the models. The regenerated load profiles have the same properties, such as power distribution and autocorrelation, as the original data.

Modelling is done with Markov models. States are defined for each customer type (Section 5.2.3). A first Markov model describes the behaviour of the customer type, electricity usage during the various days of the week are namely correlated, see the autocorrelation plot (Figure 7.3). The second Markov model adds variation to this behaviour. A profile is generated by first creating a sequence of states describing the behaviour, then used by the variation model to create the electricity demand (in states) during multiple days. The states are converted into electrical power to finish the profile. A behaviour model and seven variation models (a model for each day of the week) are built for each quarter of the year.

### 7.2.1 States

Markov chains describe transitions between discrete states (Section 3.3). Transitions are discrete in time. Load profiles are also discrete in time, but continuous in the electrical power domain. Each state of a Markov chain has to describe an interval in electrical power in order to model a load profile. To determine the borders of the states in terms of power, knowledge of its distribution is required.

The distribution of the electrical power describes how likely an electrical power value is. A curve fit through the histogram of the electrical power parametrizes the distribution. A cumulative histogram allows for a cumulative distribution fit. Cumulative distributions make it possible to divide the distribution in multiple equally likely parts. Each part has its own boundaries and is considered to be a state.

The electrical power distribution differs between the various customer types. An electrical power distribution has to be made for each customer type. The quality of a histogram depends on the number of observations to create it. The quality of a histogram can be scaled-up by using fuzzy histograms [146]. The histograms of a customer type are made fuzzy by using the weights proposed in Section 5.2.3. The power distribution histogram of one customer is weighted according to the probability the load profile is part of the customer type. The normalised sum of the weighted histograms represents the histogram of the customer type.

Histograms with bin sizes of 0.01 kW are built with both relaxed and corrected models (Section 5.2.3). The Markov chains are trained for both to make a comparison possible. Only the distribution fit with the relaxed weights is explained in the remainder of this section as the principles of fitting are the same.

Weibull distributions have the best results in power and total electricity demand curve fits (Section 5.1.1). First, an attempt is made to fit one curve though the cumulative fuzzy distribution of a customer type. The fit for the average day-consumer on the relaxed cumulative histogram is represented as ‘single fit’ (Figure 7.2).

The fit is not good enough to work with. To improve parametrisation, a double curve fit is done on the cumulative distribution. A first curve is fitted onto the first 65 % of the data, a second one on the tail, i.e. the remaining 35 % of the data. After 65 %, the slope of the down trend of the probability density curve flattens, requiring another fit to compensate for the underestimation of the tail. The parameters from the cumulative distribution fits are used

Table 7.5: Evaluation of single and joint fit for average day-consumer

| | data | single fit | joint fit |
|---|---|---|---|
| average | 0.45 kW | 0.41 kW | 0.44 kW |
| median | 0.26 kW | 0.22 kW | 0.29 kW |
| RMSE | 0 | 0.0021 | 0.00086 |

to create the non-cumulative distribution of both fits. The first 65 % of the likelihoods (non-cumulative) of the first fit and the last 35 % of the second together represent the distribution of the joint fit (after normalisation). The normalisation is done by dividing the likelihoods by the sum of the likelihoods.

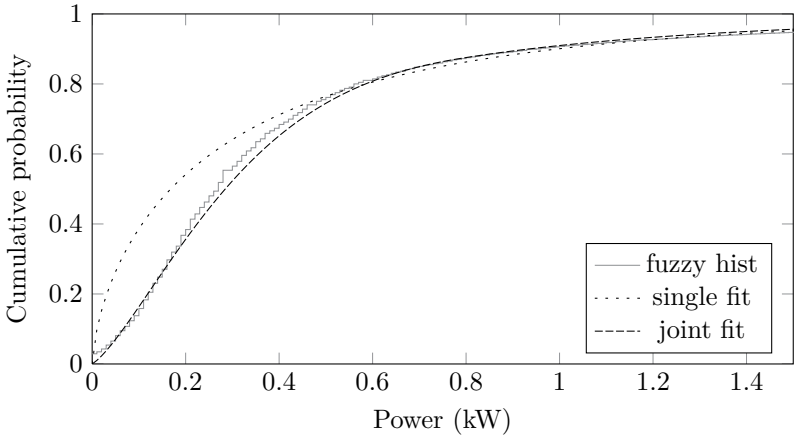
The 'joint fit' fits the data better than the 'single fit' (Figure 7.2). The difference for the average day-consumer is presented in Table 7.5. All values are calculated starting from the second bin as the first value of the 'single fit' is infinite. The average and the median power in the 'joint fit' case are closer to the data and the root mean square error (RMSE) is also lower.

The cumulative distributions of the 'joint fits' each are divided into 10 equally probable states. The state boundaries are set at the electrical power which coincides with 10 % increases of the cumulative probability. Given the skewness of the probability distributions, frequently occurring low electrical powers are modelled with more states, i.e. more detail, while lesser frequent high powers will have fewer states, i.e. less detail.

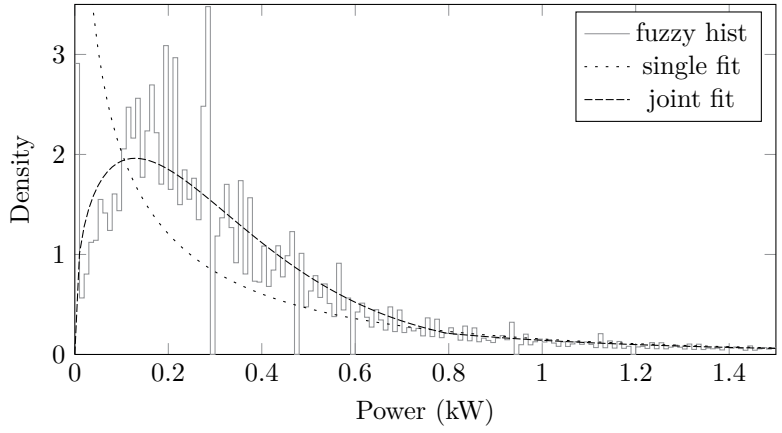
The choice for 10 states per customer type is a trade-off between level of detail, size of the transition matrix and risk of overfitting. More states result in more detailed models. However, more states also mean fewer data per state and per transition between states, i.e. a higher risk of overfitting the data. The Markov chains used to model the load profiles are second order ones, having transition matrices with size $S^3$, where $S$ is the number of states. A high number of states would result in large, sparse matrices.

### 7.2.2 Transitions

The Markov chains to model electricity demand are second order non-homogeneous ones. Non-homogeneous refers to the transitions changing each time step. The reasoning behind the non-homogeneousness is that the probability of a certain electrical power demand changes during the day. For example for the average day-consumer, it is more likely that a high power is demanded at noon, compared to 3 am (Figure 7.6, original).




(a) Single and joint curve fit of cumulative probability




(b) Single and joint curve fit of probability

Figure 7.2: Cumulative and normal probability fit of average power per fifteen minutes for the average day-consumer using relaxed weights

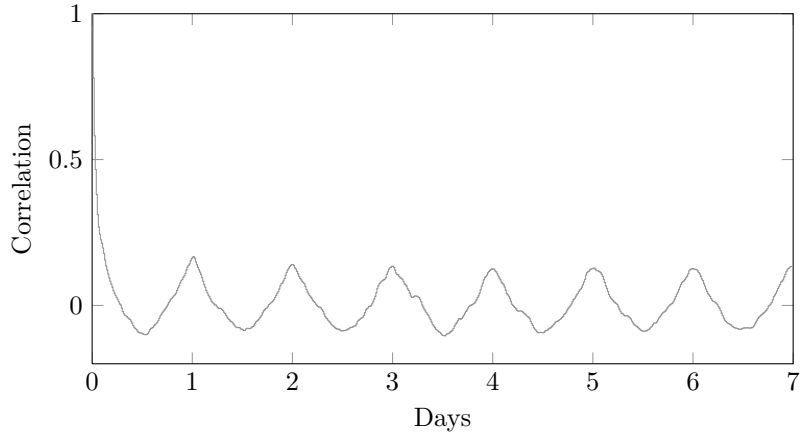



Figure 7.3: Autocorrelation of one randomly selected load profile of one year length

Second order Markov chains have a memory of size one, i.e. one value is stored: not only the current state is used to determine the next state, but also the previous one. For a randomly selected profile, the autocorrelation plot over a full year is calculated (Figure 7.3). The correlation between the current demand and the demand for the next fifteen minutes is 0.78, between the previous and the next fifteen minutes 0.58, still being high. The autocorrelation explains the need for a second order Markov chain: the electricity demand in a household doesn't change fast enough between fifteen minute measurements to work without a memory function.

Two types of Markov chains are trained: to build the behaviour of a customer and to add variation to the resulting behaviour. Only one Markov chain did not suffice to recreate load profiles: there isn't any information about the difference in electricity demand during the different days of the week and the load profiles are too smoothed.

The difference between regular Markov chains and the Markov chains here is the way the transitions are calculated. Instead of counting the number of transitions and normalising them, the probability of belonging to a certain customer type is used in the construction of the transition matrices for that customer type. A customer that is more likely to belong to a given type has a larger impact on these transition matrices. Also, more data is used to construct the transition matrices because customers with a low probability of belonging to a customer type still impact the transition matrix.

### Behaviour chain

The autocorrelation (Figure 7.3) shows that the behaviour of one day is correlated to that of other days. In this case, the correlation between a day and the next is 0.167. The correlation stays about the same for other succeeding days. The behaviour of a customer can thus be built from one day of the week. Weekdays are very similar, Saturdays and Sundays are slightly different with respect to electricity demand. Monday is a weekday and hence selected as day to train upon.

The behaviour Markov chain is built as follows. $X^b(t)$ is a stochastic process with 96 time steps, a time step $t$ for every fifteen minutes of a day. The behaviour is modelled by $Pr^b$, the probability of going to state $i$ in the current time step $t$, given previous state $j$ and the state before $k$,

$$Pr^b\{X^b(t) = i | X^b(t-1) = j, X^b(t-2) = k\} = P^b_{i,j,k}(t, t-1, t-2) \quad (7.5)$$

Each time step $t$ has its own transition matrix $P^b_{i,j,k}(t, t-1, t-2)$. Each customer type has its own behaviour Markov chain. For each quarter of the year and each customer type, a behaviour Markov chain is built. The use of quarters makes it possible to include seasonal effects (Section 2.2).

### Variation chain

A variation Markov chain is built for each day of the week of each quarter of the year for each customer type. The purpose of the variation chains is to include day-specific information in the sequence of resulting states. Electricity demand on Sunday for example is different from electricity demand on Mondays, although there is some relation between them. The difference between the average week- and the weekend day of the average day-consumers is visualised in the 'original' curves of Figure 7.6.

Again, $X^v(t)$ is a stochastic process with 96 time steps. The variation is modelled by $Pr^v$, which states that the probability of the current state $i$ depends on the state $j$ at the same time step $t$ of the behaviour model $X^b$ and on the previous state $k$. Each time step has its own transition matrix $P^v_{i,j,k}(t, t, t-1)$.

$$Pr^v\{X^v(t) = i | X^b(t) = j, X^v(t-1) = k\} = P^v_{i,j,k}(t, t, t-1) \quad (7.6)$$

### 7.2.3 Data generation

Load profiles are generated by taking samples of one step of the Markov chains. First, states are sampled from the behaviour Markov chain, resulting in a sequence of behaviour states. Secondly, the behaviour states are used as input for the variation Markov chains. Single step samples are taken to ensure that the initial state (the behaviour) isn't forgotten [147]. States representing demand during a specific day are sampled from the variation Markov chains. The states are translated into electrical power by taking a random sample of the interval of the state and passing it to the inverse of the joint distribution fit.

The process is described more formally in 4 steps:

1. Select a cluster to generate a profile for.
2. Create the general behaviour of the customer.
    - Randomly select two start states ($X^b(t_0 - 1)$ and $X^b(t_0)$) according to their probability.
    - Take one-step samples from the the behaviour Markov chain until all behaviour states ($X^b(t)$) are generated.
3. Create the behaviour of the customer on different days.
    - Randomly select a start state for the day ($X^v(t_0)$), according to the probability of the state.
    - Use the start state of the day ($X^v(t_0)$) and the start state of the model ($X^b(t_0)$) to create the next state ($X^v(t_1)$).
    - Take one-step samples from the detail Markov model of that day until the detailed behaviour of the customer during that day is generated.
    - Repeat for all days of the week and repeat multiple times if more than one week is needed.
4. Convert the states into power.
    - Randomly sample a value from the power distribution within the limits of the state.
    - Repeat for each state until the load profile is generated.

### 7.2.4 Validation

The results are validated by comparing generated with measured profiles. The original clustering weights, not the relaxed or the corrected one, are used for the

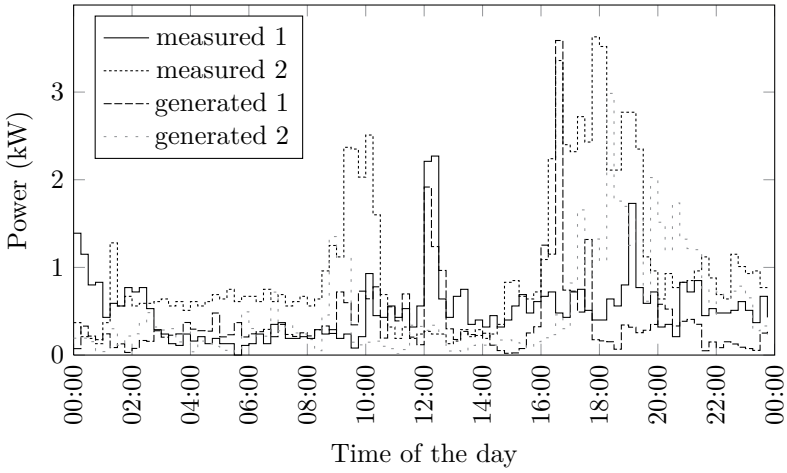



Figure 7.4: A Thursday in the first quarter of two measured and two generated profiles from the average day consumer group

measured profiles (Section 5.2). The focus of the validation is on four aspects of the results: average power, power distribution, shape of the cluster centre and autocorrelation. However, to make sure that the individual profiles are correct as well, they are compared first.

### Individual profiles

Two measured and two generated profiles from the average day consumer are picked to compare against each other (Figure 7.4). The profiles depicted show the first Thursday of the first quarter. The trend of the profiles is very comparable, showing that the model is able to generate individual profiles.

### Average power

Average power is considered to be a proxy of the total demand. The representativeness of the magnitude of the electricity demand is tested with the average power.

Table 7.6 shows the average power over all generated profiles for every customer type, calculated from the original profiles using the original cluster centre: the average of the average week per quarter per customer type is calculated.

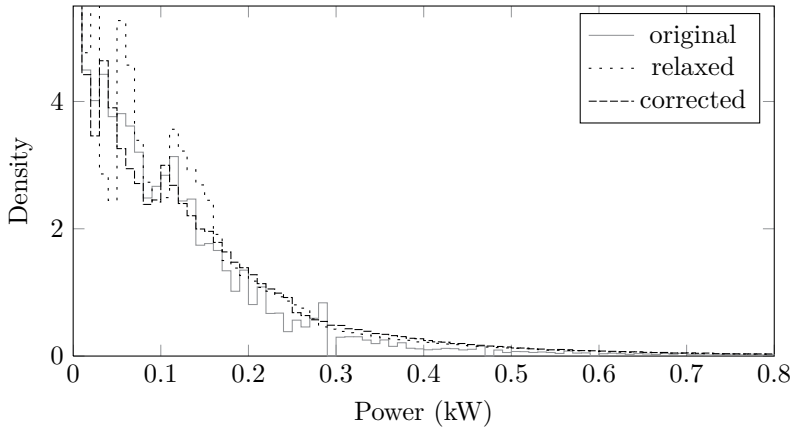



Figure 7.5: Probability mass function of the average power per fifteen minutes of the generated load profiles for the small day consumers

The average power of the various customer types are closer to the real average power when the demand of the customer type is lower. The power distribution is the reason for the lower average power, as explained later.

The corrected model performs slightly better than the relaxed except for the large business consumer where the average power completely misses the average power and relatively large business consumer where no load profiles could be generated.

### Power distribution

An electrical power distribution shows how the electrical power is distributed over all time steps and all profiles of a given customer type. Similar power distributions means that electrical power values are equally probable between sets. A comparison of the electrical power distribution of the original data and the generated data allows for the validation of the electrical power values of the generated data.

A detailed view on the electrical power distributions is shown in Figure 7.5. The distribution of the electrical power for the generated load profiles of the small day consumer is plotted against the measured data. Both the distributions of the relaxed and the corrected model follow the original distribution well, because the fuzzy histograms are very similar to the original. The technique is able to reproduce the fuzzy histograms.

Table 7.6: Average power [W] of the cluster centres and data generated from the Markov chains

| Type | | Q1 | | | Q2 | | | Q3 | | | Q4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | data | rel. | corr. | data | rel. | corr. | data | rel. | corr. | data | rel. | corr. |
| Day | small | 101 | 109 | 99 | 82 | 100 | 90 | 81 | 104 | 90 | 104 | 121 | 106 |
| | relatively small | 314 | 289 | 286 | 265 | 253 | 245 | 257 | 253 | 245 | 317 | 303 | 295 |
| | average | 534 | 434 | 449 | 444 | 374 | 385 | 433 | 372 | 378 | 538 | 457 | 470 |
| | relatively large | 848 | 583 | 649 | 684 | 491 | 532 | 647 | 478 | 519 | 848 | 608 | 668 |
| | large | 1550 | 814 | 1044 | 1204 | 639 | 804 | 1111 | 635 | 776 | 1440 | 842 | 1038 |
| Night | average | 1332 | 1421 | 2147 | 389 | 483 | 490 | 122 | 345 | 358 | 987 | 1436 | 2069 |
| | large | 1546 | 494 | 1165 | 648 | 376 | 446 | 412 | 376 | 417 | 1386 | 499 | 1217 |
| Business | average | 3781 | 1177 | 2088 | 2965 | 809 | 1503 | 2675 | 782 | 1476 | 3523 | 1240 | 1435 |
| | relatively large | 7871 | 1845 | / | 8120 | 1299 | / | 7549 | 1152 | / | 8430 | 1711 | / |
| | large | 20748 | 435 | 101 | 21538 | 286 | 102 | 22315 | 379 | 102 | 21964 | 220 | 192 |

The fuzzy histograms influenced by other clusters (Figure 5.2) differ from the original. The comparison is therefore done on boxplots. Boxplots are able to represent the distribution in a simple and convenient way. The boxplots are built in the same way as in Section 5.2.3 (Figure 5.3), the middle bar represents the power distribution of the original data, the left plot shows the results from the relaxed and the right the corrected model. The minimum and the maximum value are the $5^{th}$ and the $95^{th}$ percentile, the box boundaries are the $25^{th}$ and the $75^{th}$ percentile and the line in the box is the median. The boxplots of the various customer types are presented in Figure 7.6.

The distributions show the same trend as the average power and in the boxplots of the weighted cluster data (Figure 5.3). The boxplots of the generated and the original data of the customer types with low demand are very similar. However, the higher the demand of a customer type, the more the median and the tail of the distribution of the generated data differs from the original data. The reason is the same as for the boxplots of relaxed and the corrected cluster weights (Section 5.2.3): the distributions tend to regress to the mean. The corrected model performs in general slightly better than the relaxed model, just as in Section 5.2.3.

### Shape of cluster centres

The load curve of a customer type shows the demand of the average customer within that cluster. The timing of the demand of the generated load profiles is checked with the load curve of the customer type.

The load curves of the average week and weekend day of the second quarter of the year of the average day consumer are shown in Figure 7.7. The shape of the load curves generated by both the relaxed and the corrected model are similar to the original load curves, but the power demand is lower. The load curve of the relaxed model follows the timing of power rise in the morning during weekends better compared to the corrected. The changing probabilities of the transitions ensure higher electricity demand during the correct periods of the day for both.

The load curves of the other customer types give similar results, expect for the ones where the electrical power distribution and the average power differ from the original data.

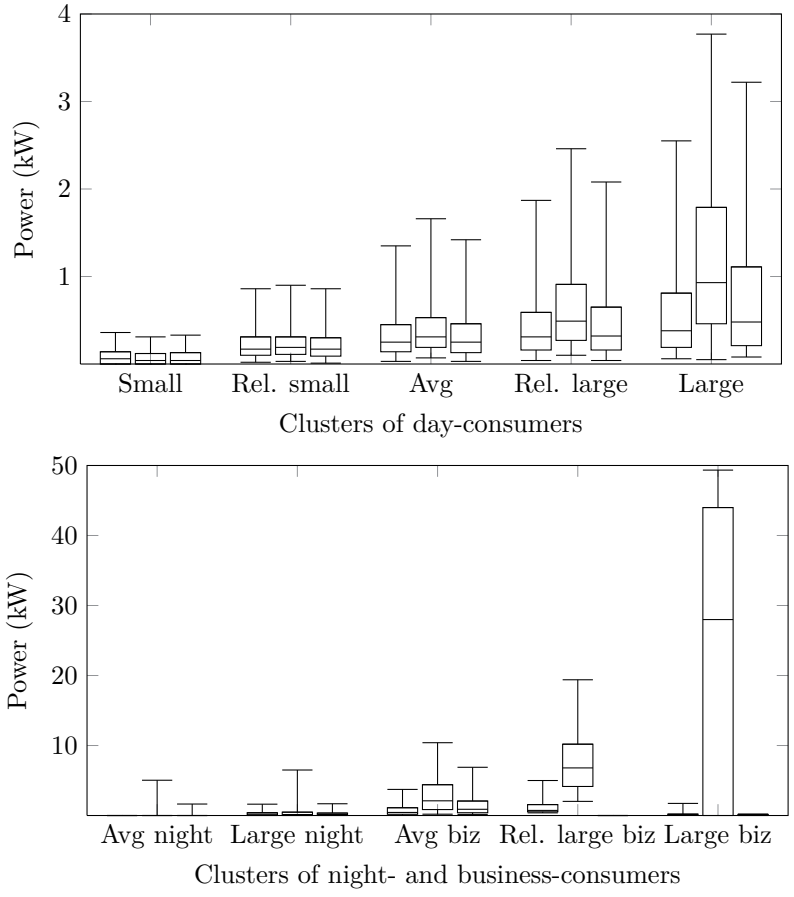
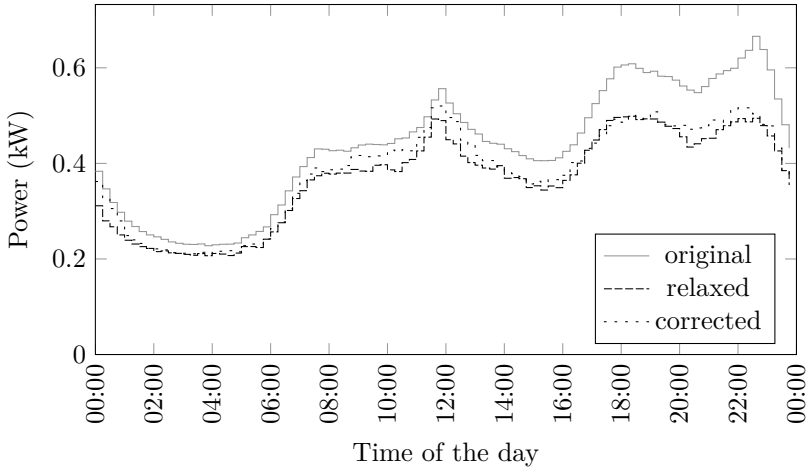
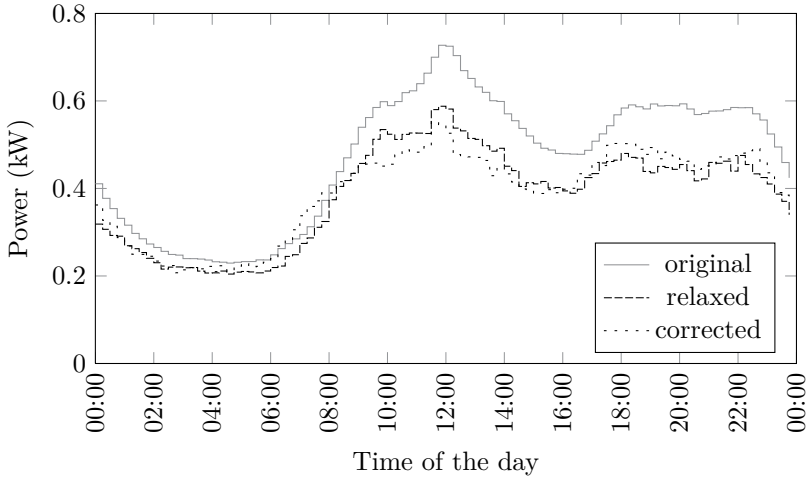



Figure 7.6: Distribution of electrical power of the original (centre) data and the generated relaxed (left) and corrected (right) data.




(a) Average weekday




(b) Average weekend day

Figure 7.7: Average week and weekend day of the average day consumer in the second quarter of the year

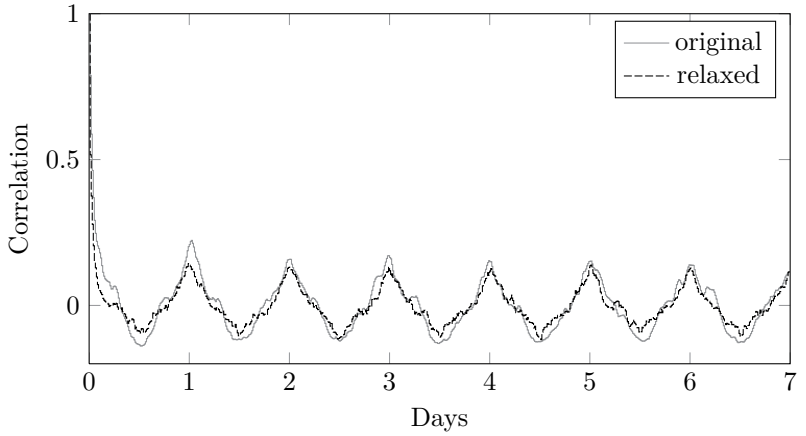



Figure 7.8: 13 week autocorrelation of one randomly selected load profile and one generated profile

#### Autocorrelation

Previous top down load data approaches are unable to reproduce the autocorrelation common to load profiles, as explained in Section 4.3.2. The autocorrelation plot of a randomly selected profile from the original and the generated data set using the relaxed model are shown in Figure 7.8. The autocorrelation is taken from a 13 week profile. The figure shows that autocorrelation is reproduced. The same results are found with the corrected model.

## 7.3 Appliance profile generation

The models describing demand at the connection point give no information on the impact of shifting appliances. To simulate those, models are required, describing the use and electricity demand of wet appliances.

The project's dataset 1 (Section 2.5.1) is the input of the appliance models. Within the appliances' measurements, cycles are detected by the algorithms (Section 6.2). The settings are estimated for each detected cycle with the help of the slightly modified settings estimation algorithms of Section 6.2.

The detected cycles and the corresponding settings are parametrised with the help of distributions and start curves, curves presenting the likelihood of starting

the appliance. An appliance model consists of the distributions and the start curves. Each customer type has its own model.

By parametrising start curves and settings distributions, electricity demand is generalised. Sampling results in the reuse of a (large) set of load cycles, while models are able to generate unseen load cycles, depending on the distribution of the parameters. The generalisations does not hold for the dishwasher, because of the limited parametrisation possibilities. Data generation is also privacy friendly: generated load cycles are not coupled to a household.

### 7.3.1 Approach

The appliance models consist of start curves and probability distributions over the settings of the appliance. Each customer type has its own appliance models and each model is valid for one quarter of the year. The models are built in four steps,

- Cycles of the appliances are detected from sub-metering data.
- Appliance settings are determined based on the load cycle found.
- Start timings of all appliances are aggregated into a start curve; the impact of a start on the start curve depends on the cluster membership.
- Settings are aggregated into a histogram; a curve fit ensures settings' parametrisation.

The considered appliances are washing machines, tumble dryers and dishwashers. Only a reduced set of settings is taken into account because of the wide variety of settings for each of these appliances. The goal is to regenerate load profiles, not to correctly estimate the distribution of settings.

The models are validated by comparing them with load curves of Section 5.3 and by comparing the start curves with those in the literature [47].

### 7.3.2 Washing machine

Section 6.2.1 describes the operation principle of a washing machine. The cycle detection algorithm remains the same. However, the detection of the settings is adjusted to reduce the number of parameters to model washing machine usage.

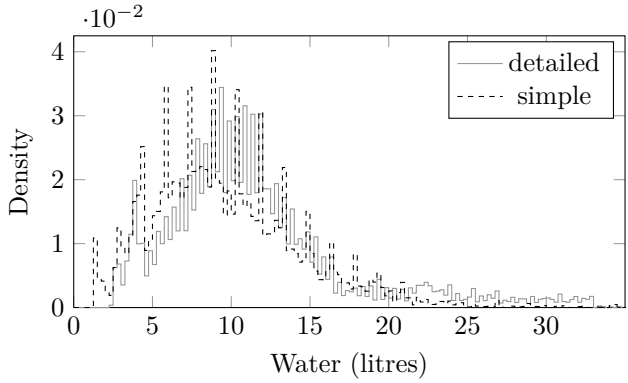



Figure 7.9: Distribution of water demand when detecting settings with detailed and simplified algorithm

### Detection of settings

Section 6.2.1 describes an algorithm to detect the settings of a washing machine. The settings are related to the water demand and the temperature setting. In this section, the algorithm is simplified: laundry weight in stead of water demand is used as modelling parameter, i.e. only cottons programmes are considered. The reasoning behind the simplification is the distribution of the water demand: there is an uplift in the tail, which makes it harder to fit a distribution. The downside of the simplification is the loss of programme information. The extra water demand is however compensated by a higher temperature setting.

Figure 7.9 illustrates the distribution of the water demand for the detailed settings estimation of Section 6.2.1 and the simplified version for the average day-consumer. The data source for the plot is the project's dataset 1 (Section 2.5.1). The fuzzy histograms are created with the cluster membership weights as explained in Sections 5.3.

The load profiles of the project's dataset 1 are used to compare the consequences of the simplification of the algorithm. The same randomly selected load profiles of Section 6.2.1 are used. Table 7.7 shows how the original detailed algorithm compares to the simplified algorithm. Although the performance of the simplified version in recalculating the total demand is better in some cases, the root mean square deviation RMSD (6.3) of the simplified algorithm indicates a higher error for each individual cycle.

Table 7.7: Consequences of the simple settings algorithm

| dw nr. # | cycles # | detected [kWh] | calc detail [kWh] | calc simple [kWh] | RMSD detail | RMSD simple |
|---|---|---|---|---|---|---|
| 5 | 242 | 126.28 | 128.36 | 125.56 | 0.0317 | 0.0465 |
| 8 | 9 | 8.16 | 8.20 | 8.50 | 0.0382 | 0.0619 |
| 13 | 60 | 46.19 | 46.33 | 46.18 | 0.0396 | 0.0450 |
| 25 | 10 | 8.01 | 7.99 | 7.91 | 0.0345 | 0.0466 |
| 28 | 184 | 131.95 | 135.79 | 133.57 | 0.0471 | 0.0515 |
| 32 | 596 | 323.91 | 321.51 | 310.26 | 0.0423 | 0.0700 |
| 42 | 172 | 163.29 | 166.63 | 165.68 | 0.0380 | 0.0468 |
| 45 | 343 | 223.50 | 220.40 | 220.35 | 0.0302 | 0.0352 |
| 66 | 526 | 428.04 | 442.07 | 440.16 | 0.0452 | 0.0544 |
| 67 | 140 | 56.49 | 55.72 | 52.68 | 0.0283 | 0.0505 |

### Settings probabilities

The probabilities of the laundry weight and temperature setting need to be determined for each customer type, the data upscaling technique (Section 5.3.2) being the basis for the construction of the distributions. The same limitations related to the data hold as for the demand description of washing machines (Table 5.7) holds: there is only sufficient data for the relatively small, the average, the relatively large and the large day-consumer, together with the large night and the average business consumer.

The average temperature settings of the considered consumer groups are comparable: the lowest average temperature is 53.3 ℃ (average day-consumer), the highest is 56.3 ℃ (relatively small day-consumer). Therefore, the weighted average of the temperature settings is used for all customer types. The weighing is done base on the relaxed customer type probabilities (Section 5.2.3). The probabilities of the various temperature settings are displayed in Table 7.8. The probabilities described in the literature [48] are shown as well, cold (no temperature) having a probability of 1.3 %. The detailed model approximates the probabilities of the literature best. Higher temperatures have, as expected, higher probabilities in the simplified settings detection algorithm.

The distributions of the laundry weights of the considered customer groups are also similar, as shown in Figure 7.10. A new distribution is composed of the weighted average of the weight distributions of the customer types. Weighing is again done based on the relaxed customer type probabilities. A Weibull distribution (5.1) is fitted through the composed distribution with shape parameter $k$ of 2.1504 and scale parameter $\lambda$ of 4.8576. The $\chi^2$-error (5.4) is

Table 7.8: Distribution of temperature settings detected by the detailed and the simplified algorithm.

| Temperature setting | 30 °C | 40 °C | 60 °C | 90 °C |
|---|---|---|---|---|
| Simple | 17.2 % | 28.5 % | 29.7 % | 24.6 % |
| Detailed | 24.8 % | 30.7 % | 34.9 % | 9.6 % |
| Literature | 23.3 % | 40.8 % | 30.6 % | 4.0 % |

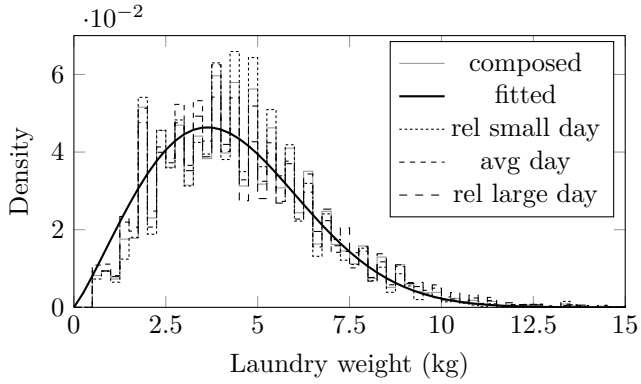



Figure 7.10: Washing machine weight distribution for various customer types and the fit through the composed distribution

2.66, which coincides with a high confidence.

The cycle starts are aggregated into start curves. The start curves depict the likelihood of starting an appliance given the time of the day. A washing machine start curve is created for every household with a washing machine in the project's dataset 1. The weighted average of the start curves results in the start curves of the different customer groups. The washing machine's start curve of the average day consumer is depicted in Figure 7.11.

The sum of all likelihood values of a start curve results in the average number of starts per week. Relatively small day-consumers operate their washing machine 4.4 and average day-consumers 5.2 times per week on average. Relatively large and large-day consumers start theirs on average respectively 5.4 and 5.5 times per week. The frequencies for large night and average business-consumers are 5.2 and 5.3 times per week. Except for the relatively small day-consumer, the operation frequency of the customers is similar.

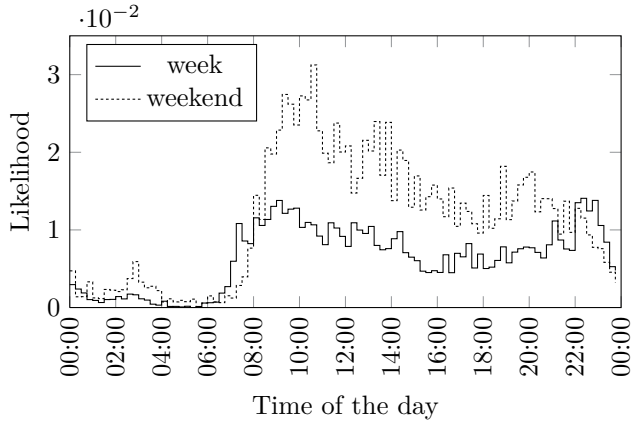



Figure 7.11: Washing machine's start curve of the average day consumer

#### Validation

The model, consisting of start curves, temperature probabilities and a laundry weight distribution, is validated by regenerating the appliance load demand curve. For each customer type, 1000 one week load profiles are generated and aggregated into a load curve. The measured washing machine load demand curves are defined in Section 5.3.2.

Figure 7.12 shows the measured and the generated washing machine load curves of the average day-consumer. The shape of both is highly comparable, but the peak power demand is lower. The same holds for other the customer types.

Table 7.9 compares the average power during week and weekend days for the measured and the generated load curves. The lower peak power in the generated load curves explains the lower average power for all customer types.

### 7.3.3 Tumble dryer

The operating principle of a tumble dryer is explained in Section 6.2.3. The cycle and the settings detection algorithms remain the same. However, for consistency with the washing machine, laundry weight instead of residual moisture is reported. The laundry weight is calculated from the residual moisture by assuming a moisture level of 55 %, coinciding with a washing machine's dry spinning speed of 1100 rpm.

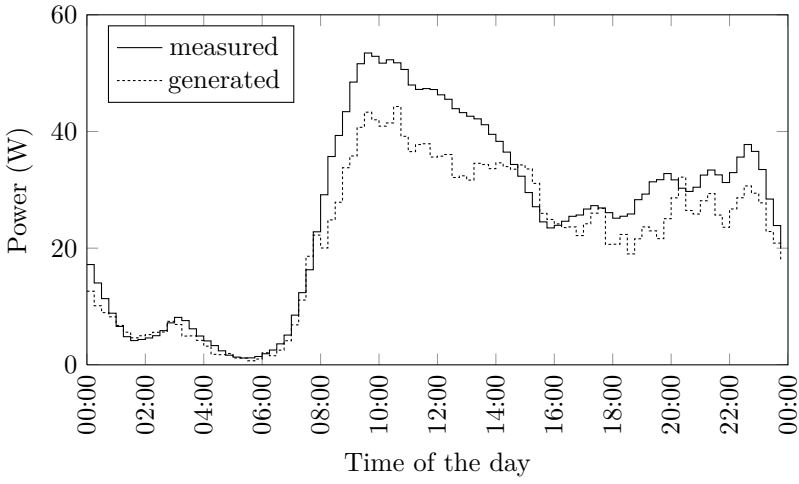



Figure 7.12: Measured and calculated washing machine load curves of the average day-consumer

Table 7.9: Average washing machine's power during weekdays and weekends for the measured and the generated load curves of the different customer types

| Type | Sub-type | Measured [W] | | Generated [W] | |
|---|---|---|---|---|---|
| | | weekday | weekend | weekday | weekend |
| Day | relatively small | 19.1 | 28.3 | 16.0 | 24.1 |
| | average | 20.7 | 38.9 | 18.1 | 30.9 |
| | relatively large | 22.2 | 43.0 | 18.1 | 35.6 |
| | large | 23.1 | 42.3 | 18.3 | 33.5 |
| Night | large | 20.1 | 39.4 | 16.8 | 31.8 |
| Business | average | 22.6 | 41.2 | 17.7 | 31.8 |

Table 7.10: Distribution of detected heating resistor powers

| Resistor power [kW] | 0.5 | 1.0 | 1.5 | 2.0 |
|---|---|---|---|---|
| Probability [%] | 11.1 | 30.2 | 28.8 | 29.9 |

### Settings probabilities

Two settings' parameters are considered for the tumble dryer model: power of the heating resistor and laundry weight. The data source for modelling is the project's dataset 1 (Section 2.5.1). The distributions and the start curves are built using the scaling up technique of Section 5.3.2. The limitation of Table 5.7 holds: only the relatively small, the average, the relatively large and the large day-consumer groups, together with the large night and the average business-consumer groups have sufficient data.

The resistor power settings of the various customer types are comparable: the minimum average power is 1.34 kW (relatively large day-consumer), the maximum average power is 1.37 kW (large night-consumer). The weighted probabilities are hence used for all customer types. The weighing is done based on the the relaxed customer type probabilities (Section 5.2.3). Table 7.10 shows the distribution.

The distribution of laundry weight placed in the tumble dryers is also similar amongst the customer groups. Figure 7.13 depicts the weight distributions of the relatively small, average and relatively large-customers. The composed distribution is the weighted average of the distributions of the various customer groups. A Weibull curve is fitted through the composed distribution. The value of the shape parameter $k$ is found to be 1.3556 and the scale parameter $\lambda$ has 2.8762 as value. The $\chi^2$ error is 0.46.

The cycle starts of the individual users are converted into start curves for the customer groups in the same way as done for washing machines. The sum of the likelihoods results again in the average number of starts per week. Relatively small day-consumers are found to operate their tumble dryer on average 3.22 times per week, average day-consumers 3.64, relatively large 3.68 and large day-consumers 3.33 times per week. The average number of starts per week for the large night and average business-consumer are 3.49 and 2.56.

### Validation

The start curves, the resistor power probabilities and the laundry weight distributions jointly represent the tumble dryer model. Models are validated

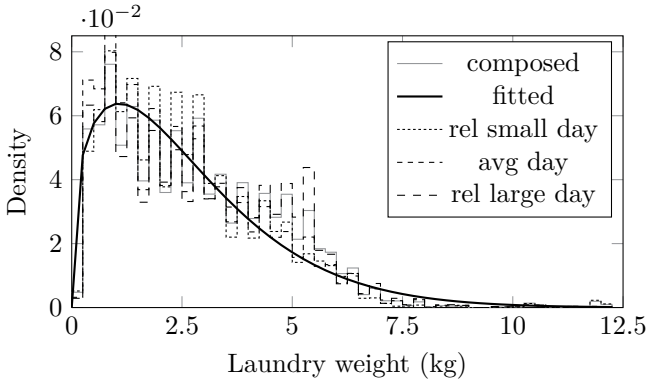



Figure 7.13: Tumble dryer weight distribution for various customer types and the fit through the composed distribution

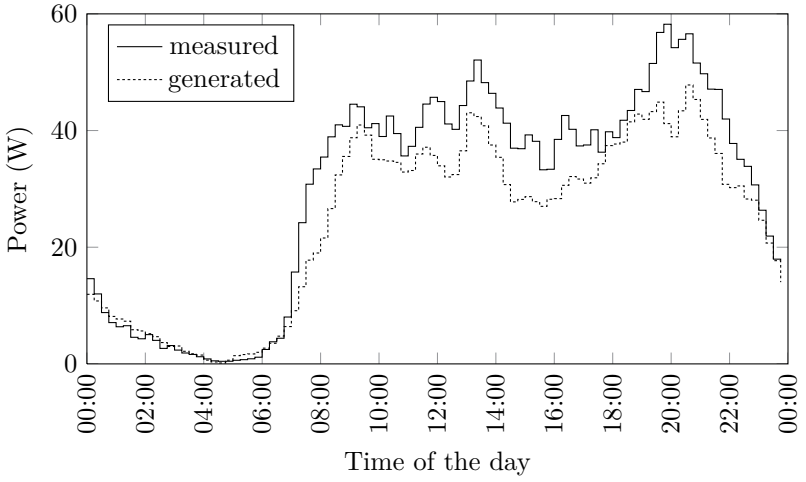



Figure 7.14: Measured and calculated tumble dryer load curves of the average day-consumer

by regenerating the tumble dryer load curve. 1000 one week load profiles are generated for each customer type. The profiles are aggregated into a load curve and compared to the original measured tumble dryer load curves (Section 5.3.2).

Measured and generated load curve of the average day-consumer are depicted in Figure 7.14. The load curve shape is very similar, however the peak demand is lower. The same pattern is true for the other customer types.

Table 7.11: Average tumble dryer's power during weekdays and weekends for the measured and the generated load curves of the different customer types

| Type | Sub-type | Measured [W] | | Generated [W] | |
|---|---|---|---|---|---|
| | | weekday | weekend | weekday | weekend |
| Day | relatively small | 22.1 | 32.9 | 19.5 | 27.4 |
| | average | 25.0 | 40.3 | 21.6 | 32.1 |
| | relatively large | 26.6 | 52.9 | 21.1 | 36.4 |
| | large | 26.5 | 52.7 | 19.5 | 34.2 |
| Night | large | 20.9 | 47.5 | 19.3 | 37.4 |
| Business | average | 25.7 | 51.7 | 14.4 | 27.8 |

The average power demand during weekdays and weekends for the measured and the generated load curves are shown in Table 7.11. The lower peak demand is also visible in the lower average power for all customer types.

### 7.3.4 Dishwasher

Section 6.2.2 describes the operating principle of a dishwasher. The cycle and the settings detection detection algorithms are the same as the original, except for the plate settings. Only 1 of the 50 dishwashers has a plate setting different from 12, namely 9. Plate setting is therefore dropped from the parameters.

#### Settings probabilities

The temperature settings need to be determined for each customer type. The project's dataset 1 (Section 2.5.1) is used in combination with the scaling up technique (Section 5.3.2). The limitations of Table 5.7 hold: only relatively small, average, relatively large and large day-consumer groups, together with large night and the average business-consumer groups have sufficient data.

The temperature settings are comparable for all customer types. The lowest average temperature amongst the considered customer groups is 46.7 °C (relatively large day-consumer), the highest is 51.6 °C. All other customer groups have an average temperature in between. The weighted average of the temperature distributions is shown in Table 7.12, together with the temperature probabilities as found in the literature [48].

The distribution differs from the literature in a preference for higher temperatures of the detection algorithm. One explanation could be that the

Table 7.12: Distribution of heating resistor powers detected.

| Temperatures [°C] | 40 | 50 | 60 | 70 |
|---|---|---|---|---|
| Probabilities [%] | 6.4 | 23.5 | 39.9 | 30.2 |
| Literature [%] | 13.9 | 36.3 | 35.6 | 14.2 |

algorithm is developed for dishwashers with a higher efficiency than the ones in the project's dataset 1.

Start curves are created from the cycle starts of the dishwashers. A start curve for each customer type is made by using the weighted average of the individual start curves. The sum of the likelihoods of a customer type's start curve results in the average number of dishwasher starts. Relatively small day-consumers start their dishwasher on average 3.23 times per week, average day-consumers 4.24, relatively large and large day-consumers respectively 4.16 and 3.93 times per week on average. For large night and average business-consumers, the operation frequency is 3.98 and 3.91 times per week on average.

### Validation

The dishwasher models, i.e. the start curves and the temperature probabilities, are validated by regenerating the appliance load curves. 1000 one week load profiles are generated for each customer type. The load profiles are aggregated into a load curve and compared to the measured load curves of the original measured profiles.

Measured and generated load curves of the average day-consumers are depicted in Figure 7.15. The shape of both are very comparable, but the peak power is lower. The other customer types have the same problem in their load curve: shapes are good, peak power is too low.

The lower peak power results into a lower average powers. Table 7.13 shows the average powers during weekdays and weekends for the various customer types. The relative difference between measured and generated is highest during weekends.

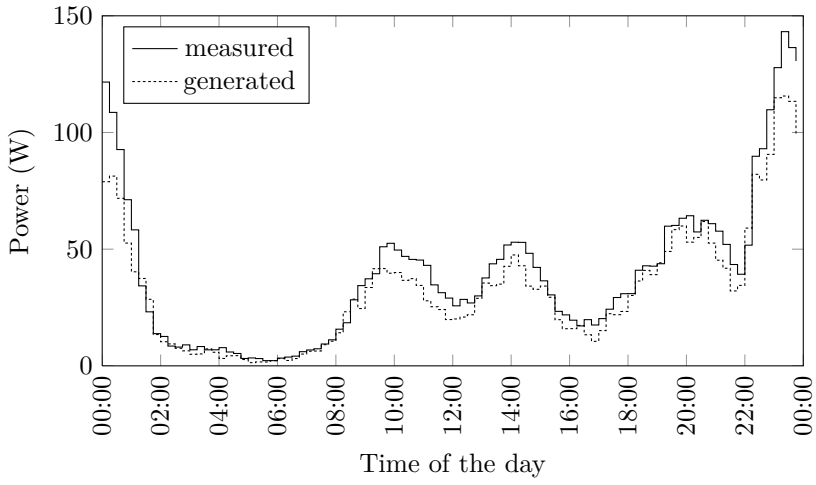



Figure 7.15: Measured and calculated dishwasher load curves of the average day-consumer

Table 7.13: Average dishwasher's power during weekdays and weekends for the measured and the generated load curves of the different customer types

| Type | Sub-type | Measured [W] | | Generated [W] | |
|---|---|---|---|---|---|
| | | week | weekend | week | weekend |
| Day | relatively small | 28.9 | 36.0 | 23.7 | 28.3 |
| | average | 36.8 | 45.2 | 31.4 | 36.24 |
| | relatively large | 35.0 | 43.7 | 29.6 | 34.3 |
| | large | 34.7 | 43.3 | 27.6 | 34.4 |
| Night | large | 32.7 | 41.7 | 29.5 | 35.2 |
| Business | average | 40.0 | 45.3 | 28.2 | 32.7 |

## 7.4 Conclusions

The selection of a representative set of customers for a field test is best done by probability sampling, because of the theoretical soundness. However, simple random sampling often suffers from the over-representation of people that want to participate. Stratified sampling, also a probability sampling technique, requires the definitions of strata, i.e. probabilities of combinations of demographic properties to select upon. The probabilities of those combinations are mostly not available. Non-probability sampling is cheaper and easier to implement, but doesn't rely on random selection.

Quota sampling is a non-probability sampling technique requiring distributions of demographic properties to sample customers. The technique is related to stratified sampling, except that only the individual distributions rather than the combined distributions of demographic properties are needed. Distributions over demographic properties are described by the government. The government is hence considered to be an expert in the definitions of the distributions. The machine learning algorithm, pointing out the demographic properties related to electricity demand, is considered to be the expert in the field of the combination of electricity consumption and demographic properties.

Sampling is executed by optimisation algorithms to eliminate the human error in quota sampling as much as possible. Both constraint programming, an artificial intelligence approach, and mathematical programming are compared. Constraint programming works with logics and requires permutations to find the optimal set. Mathematical programming does not have the permutation limitation and converges faster. The result is a set of customers or load profiles representative for the distributions.

Data generation is a solution when external factors limit the use of original data. Modelling residential load profiles makes it also possible to do Monte Carlo simulations. Load profiles are modelled using Markov chains. The distribution of the electricity demand of a customer type is first divided into states. The behaviour and the variation Markov chains work with the same states. Transitions are defined between states and transition probabilities change each time step. The relaxed customer type membership is used to build the distributions and the transitions.

To create a load profile for a customer type, states need to be sampled from the customer type's behaviour Markov chain first. The behaviour states are the input for the customer type's variation Markov chains and represent the electricity demand behaviour pattern of the household. The variation Markov chains add variation to that general behaviour. Each day of the week has its own variation chain. The result of sampling from the variation chains is a

sequence of states corresponds to the electricity demand of a household during a week, or multiple weeks. The states are converted into electrical power by passing them through the inverse distribution of the electrical power. Only one step samples are required to generate load profiles with the Markov chains. The technique is not related to Markov Chain Monte Carlo.

The profile generator is validated by generating large numbers of load profiles, aggregating them per customer type and comparing the aggregation with the aggregation of the original data. Average power, power distributions and shape of the load curves are compared. The average power of the generated profiles is lower than the average power of the original data. Correcting the relaxed customer type membership improves the result, except for one customer type where profiles could no longer be generated. The power distributions also indicate that powers are lower compared to the original load profiles. The autocorrelation of the original load profiles is found in the generated load profiles, which is a plus compared to other top-down load profile generators described in the literature. Two measured and two generated load profiles are depicted to verify that the trends are comparable.

Active demand simulations with appliances are only possible with data or models that describe the way the appliances are used and the corresponding electricity demand (Section 7.3). Cycle detection and settings estimation algorithms for wet appliances described in Chapter 6 are adapted to create start curves and settings probabilities for each appliance and each customer type. The electricity demand at the connection point of the sub-metered households (from measurements in the project's dataset 1) are used to determine the weight the corresponding appliance has in the customer type's start curves and settings probabilities. The relaxed customer type membership is used as weight.

Load cycles are randomly sampled from the start curves and the setting probabilities or the settings probability distributions, to generate an appliance's load profile. The resulting appliances' load profiles, i.e. sequence of cycles of one appliance, are validated by comparing them to the original data. The average power of the generated load profiles is lower than the average power of the original load profiles for all appliances and all customer types. The shape of the customer types' load curves resembles the original load curves.

Models of appliances are, in contrast to sampling load cycles from the measurements, better in generalising their and protect privacy better. An example of generalisation is the distribution of the weight of a washing machine: by working with a distribution, unseen weights can be sampled. The aggregation because of distribution fits also ensures privacy: the settings cannot be tracked to a customer.

# Chapter 8

# Flexibility

Flexibility is the amount of electrical load shiftable or curtailable (Section 4.4). The flexibility described in this chapter is related to wet appliances: washing machines, tumble dryers and dishwashers.

First, the potential for flexibility is estimated. The attitude of customers is linked to the ownership rate of the appliances and combined with the appliance electricity demand (Section 5.3). The result is the impact the appliances have on the total electricity demand, translated into potential, expressed in Wh per fifteen minutes.

The effect of using flexibility is tested. The question 'What happens to the demand of a number of households if appliances' starts are delayed?' is answered.

## 8.1 Potential

The share of power demand of the wet appliances compared to the overall demand determines the main part of the potential for flexibility and is called the impact $P_{imp}$ of the appliances. The impact is found by combining appliance ownership rates $\beta$ with the appliance load curves $P_{app}$ (Section 5.3) for each customer type $cl$ and each appliance $app$. By calculating the impact in each time step $t$ of the appliance load curves, an impact curve is constructed.

$$P_{imp,cl,t} = \sum_{app}^{n_{app}} \beta_{app,cl} \cdot P_{app,c,t} \tag{8.1}$$

Table 8.1: Appliance ownership rates for the various customer types

| Type | Sub-type | Ownership rate [%] | | |
|---|---|---|---|---|
| | | wash. mach. | dryer | dishwasher |
| Day | small | 85.4 | 46.7 | 26.3 |
| | relatively small | 98.1 | 69.8 | 57.4 |
| | average | 98.0 | 83.9 | 76.7 |
| | relatively large | 97.3 | 86.9 | 79.5 |
| | large | 71.9 | 60.3 | 54.1 |
| Night | average | 100. | 48.3 | 7.8 |
| | large | 100. | 87.5 | 62.8 |
| Business | average | 8.83 | 8.83 | 34.2 |
| | relatively large | 54.9 | 19.3 | 100. |

The residual demand $P_{res}$ is determined by subtracting the impact of the appliances $P_{imp}$ from the total demand $P_{tot}$.

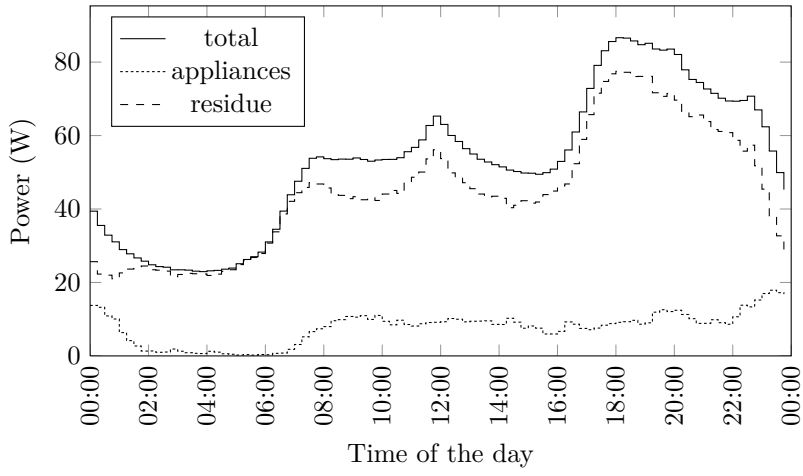
$$P_{res,cl,t} = P_{tot,cl,t} - P_{imp,cl,t} \tag{8.2}$$

The appliance ownership rates of the various customer types are shown in Table 8.1. The ownership rates are based on the answers to the survey of the Linear project (Section 2.4) and the original cluster membership (Section 5.2.2). Cluster- or customer type-memberships are therefore crisp, no relaxation, i.e. weights, have been used. There was no need to scale up the data.

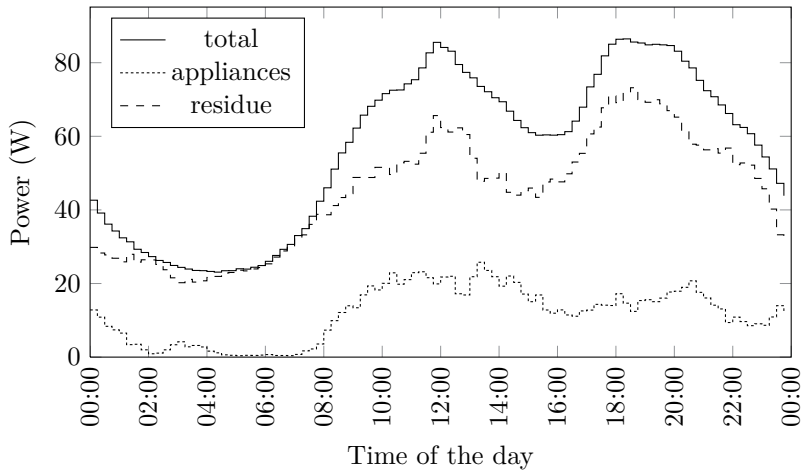
The appliances ownership rates raise with the total electricity demand: a higher electricity demand means a higher probability of owning a device. The drop in appliance ownership rates from relatively large to large day-consumers is because of the inclusion of some business-consumers in the large day-consumer group. Washing machines are more frequently owned than tumble dryers, which are on their turn more frequently owned that dishwashers.

The load curves of the total electricity demand of the average weekday and the average weekend day for the fourth quarter of the year of the average day-consumer as calculated in Section 5.2.2 is depicted in Figure 8.1. The figure also includes the joint load curves of the appliances for that period. The joint appliance load curves represent the impact of the appliances on the total electricity demand. The residue is the load curve of the total demand minus the appliance load curve.

Figure 8.1 shows the relatively modest impact of the wet appliances on the total electricity demand. The impact is almost flat during weekdays, except for




(a) week




(b) weekend

Figure 8.1: Load curves of the total demand, the demand of the appliances and the residual demand for an average day-consumer in the fourth quarter

Table 8.2: Average power of white good appliances per household in Belgium

| Type | Subtype | Impact [W] | | |
|---|---|---|---|---|
| | | mean | median | max |
| Day | relatively small | 56.6 | 55.1 | 225.8 |
| | average | 80.1 | 80.6 | 294.1 |
| | relatively large | 86.8 | 78.6 | 340.8 |
| | large | 61.1 | 51.7 | 275.6 |
| Night | large | 72.6 | 62.5 | 263.8 |
| Business | average | 27.9 | 15.4 | 256.4 |

the period between 1 a.m. and 8 a.m and the uplift in power from 10 p.m. to 1 a.m. where the influence of the tariff schemes in Belgium (Section 2.2) are visible. The impact is higher during the weekend compared to the week.

Joint appliance load curves of the average week- and weekends days for each quarter of the year have been created for the various customer types. The mean, median and the maximum value of the load curves are shown in Table 8.2. Only the customer types with sufficient appliance data are considered (Table 5.7). The mean values are a proxy for the total yearly power. A median value divides the distribution of the powers in two: the powers are higher/lower than the median value for 50 % of the time. The maximum value indicates the upper boundary. The appliance power impact is highest for average and relatively large day-consumers.

The possible attitudes of people towards the direct load control of appliances are described in Section 4.4.3. They are the basis to estimate the number of households participate in active demand. The probabilities of the various attitudes for each customer type are shown in Table 8.3. The percentages are calculated with the original cluster membership, based on the survey data (Section 2.4).

Customer types with a higher demand have in general a more positive attitude towards active demand participation. Relatively large day-consumers, night-consumers and average business-consumers have the most positive attitude.

The combination of attitude of the households and electricity demand of their appliances is the potential for active demand. Only the households with the most positive attitude are assumed to participating in active demand. The percentage of advocates per considered customer type are multiplied by the impact to find the potential of active demand per household. The numbers are also translated into energy to give insight in the energy that can be shifted on average per household (Table 8.4).

Table 8.3: Attitude of the various customer types towards active demand

| Type | Sub-type | Attitude [%] | | | |
|---|---|---|---|---|---|
| | | advocate | supporter | sceptic | refuser |
| Day | small | 27.8 | 16.7 | 26.6 | 29.0 |
| | relatively small | 28.2 | 21.6 | 36.2 | 14.0 |
| | average | 33.9 | 25.5 | 21.8 | 18.9 |
| | relatively large | 50.8 | 35.4 | 9.7 | 4.1 |
| | large | 18.5 | 48.0 | 21.5 | 12.0 |
| Night | average | 48.3 | 43.2 | 8.6 | 0.0 |
| | large | 57.9 | 29.6 | 12.5 | 0.0 |
| Business | average | 57.9 | 13.5 | 17.9 | 10.7 |
| | relatively large | 0.0 | 36.6 | 63.4 | 0.0 |

Table 8.4: Potential for active demand of white good appliances per household in Belgium

| Type | Subtype | Potential [W] | | | Potential [Wh/15min] | | |
|---|---|---|---|---|---|---|---|
| | | mean | med. | max | mean | med. | max |
| Day | relatively small | 16.0 | 15.5 | 63.7 | 4.0 | 3.9 | 15.9 |
| | average | 27.3 | 27.3 | 99.7 | 6.8 | 6.8 | 24.9 |
| | relatively large | 44.1 | 39.9 | 173.1 | 11.0 | 10.0 | 43.3 |
| | large | 11.3 | 9.6 | 51.0 | 2.8 | 2.4 | 12.8 |
| Night | large | 42.0 | 36.2 | 152.7 | 10.5 | 9.0 | 38.1 |
| Business | average | 16.2 | 8.9 | 148.5 | 4.0 | 2.2 | 37.1 |

The customer types with sufficient appliance data represent 82 % of all households, the remaining 18 % are assumed not to participate in active demand. The non-normalised weighted sum of the load curves from the customer types with sufficient data results in the potential per household in Belgium. The potential is obtained by multiplying the potential per household with the number of households in Belgium being 4.6 million. 29 % of the households (1.3 million) participate in active demand actively (every appliance start has flexibility enabled) by these assumptions, which might be an overestimation given that [47] expects 5 % of tumble dryer, 10 % of washing machine and 20 % of dishwasher cycles to be flexible. The average potential is estimated at 24 MWh per fifteen minutes, which corresponds to 96 MW. The median is 92 MW and the maximum is estimated to be 353 MW.

The 96 MW potential is low compared to the installed capacity of 19.6 GW (2011) [148], peak demand of 13.1 GW or average demand of 9.3 GW, and also

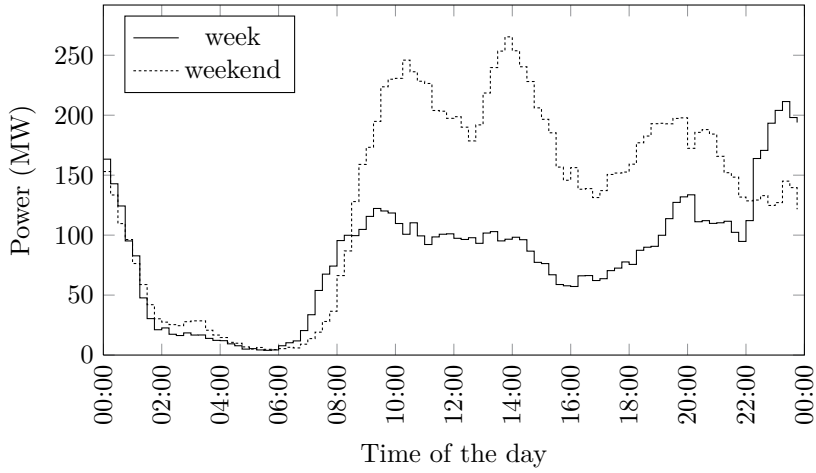



Figure 8.2: Estimated average day potential for demand response of wet appliances in Belgium

when only the residential sector is taken into account (2.3 GW) [149]. However, compared to the power reserves, the potential is not negligible. Primary reserves are 100 MW, the secondary are 137 MW. The tertiary power reserves in Belgium amount to 660 MW of which 240 MW down- and 420 MW upward. Requirements for tertiary reserves are 90 % availability for up- and 80 % for downward control [150]. The wet appliances cannot be used for primary reserves because of the response time of 30 seconds. Although a response time of less than fifteen minutes is achievable, the required availability for secondary reserves makes it hard to use appliances. The best option for the appliances seems to be operating as a last resort for balancing.

## 8.2 Effect of using flexibility

The effect on demand of using the flexibility depends on the way the flexibility is used. Various scheduling and optimisation algorithms exist (Section 4.4.4). However, algorithms that decide on the best moment to start appliances are out of the scope of the thesis. The effect of using flexibility is shown by preventing appliances from starting a new cycle until the moment they are allowed to start again.

Shifting the operation of an appliance in time is more common than splitting cycles of appliances. The prevention of starting a new cycle shows how long it

takes before flexibility takes effect. The delayed cycles are thereafter started at the same time. The joint starts illustrate the power demand of starting large numbers of appliances at once. Both give insight about potential effects of certain scheduling and optimisation algorithms.

The setup for the simulation of the effect of using flexibility consist of 100 000 participating households. The number of participating households depends the probability of the customer type and the customer type's attitudes. This probability is multiplied by the percentage of 'advocates' for the customer type. The resulting probabilities are normalised. Only the customer types with sufficient appliance data are considered (Table 5.7).

25.2 % of the 100 000 participating households are relatively small day-consumers. Average, relatively large and large day-consumers are represented by respectively 32.5 %, 27.0 % 4.9 % of the participating households. Only a small part of the households are regarded large night or average business consumers: 5.8 % and 4.6 %.

The models to simulate the wet appliances are described in Section 7.3 and are used for a weekday (a Tuesday). The appliances aren't allowed to start until 19h15. Multiple durations of preventing starts are being tested. The prevention time ranges from nil up to two hours with a step of 30 minutes. After 19h15, the appliances are allowed to start again, resulting in a large electricity demand at once. The purpose is to demonstrate the effect, not to operate a goal, therefore the exact timing and duration is not of importance.

### 8.2.1 Per appliance

The effect of using flexibility of washing machines, tumble dryers and dishwashers is simulated, using appliance models (Section 7.3) and ownership rates (Table 8.1). Keep in mind that only participating customers, i.e. advocates, are simulated to see the effect.

#### Washing machines

In total, 92 566 washing machines are simulated. The effect of delaying starts is clearly visible in the load profile (Figure 8.3). The electricity demand at release almost doubles (1.9 times) compared to the initial demand for one hour of not allowing washing machines to start. For the two hour period, the demand is almost four times (3.7 times) as high. It takes half an hour of not allowing washing machines to start before any effect is clearly visible: not every washing

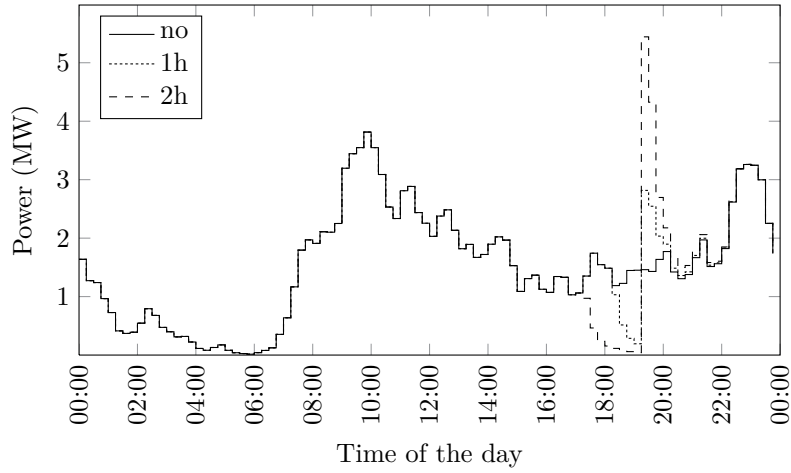



Figure 8.3: The effect on the total washing machine demand of 92566 households due to using flexibility

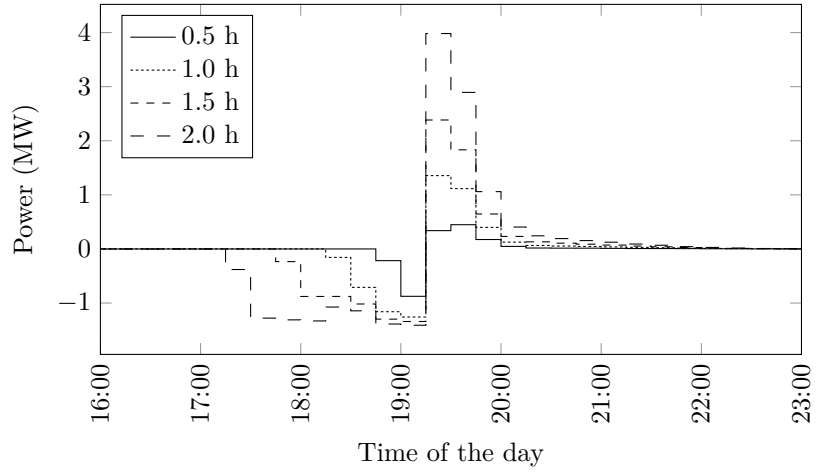



Figure 8.4: The impact of using washing machines' flexibility

machine is started at the beginning of the quarter hour and it takes about five minutes before the heating resistor is started.

Figure 8.4 shows clearer overview of the effect of using flexibility. The peak demand when washing machines are allowed to start again is relatively low (338 MW in the first fifteen minutes) when the delay is only half an hour. The

Table 8.5: Peak before and after delay of washing machines

| delay | participating [#] | power before [kW] | power after [kW] |
|---|---|---|---|
| 30' | 999 | -875 | 338 |
| 60' | 1975 | -1258 | 1355 |
| 90' | 2898 | -1341 | 2383 |
| 120' | 4276 | -1410 | 3982 |

Table 8.6: Peak before and after delay of tumble dryers

| delay | participating [#] | power before [kW] | power after [kW] |
|---|---|---|---|
| 30' | 1540 | -1582 | 507 |
| 60' | 2625 | -2339 | 1276 |
| 90' | 3691 | -2743 | 2470 |
| 120' | 4786 | -2924 | 3734 |

power demand of the delayed cycles is responsible for the relatively low demand: washing machines take time to finish, the demand in the second and the third fifteen minutes isn't demanded but postponed. Washing machines do not often operate their heating resistor longer than half an hour, explaining why the power demand is much higher for one hour and more of delay time.

The number of washing machines taking part in 30, 60, 90 and 120 minute delays are shown in Table 8.5 together with the impact before and after the delay.

### Tumble dryers

The ownership rate of tumble dryers is lower compared to washing machines: 76 747 tumble dryers are simulated in total. The impact of using the flexibility is depicted in Figure 8.5. It takes longer, compared to washing machines, before the full electricity demand of tumble dryers is gone when new tumble dryers aren't allowed to start: the difference in negative power just before 19h15 is relatively high between the various delay durations. It also takes longer before the effect of starting tumble dryers together is completely worn out.

The peak demand is relatively low (507 kW in the first fifteen minutes) when the delay is half an hour. The reason again is the power demand of the delayed cycles: cycles take longer to finish than fifteen minutes or half an hour.

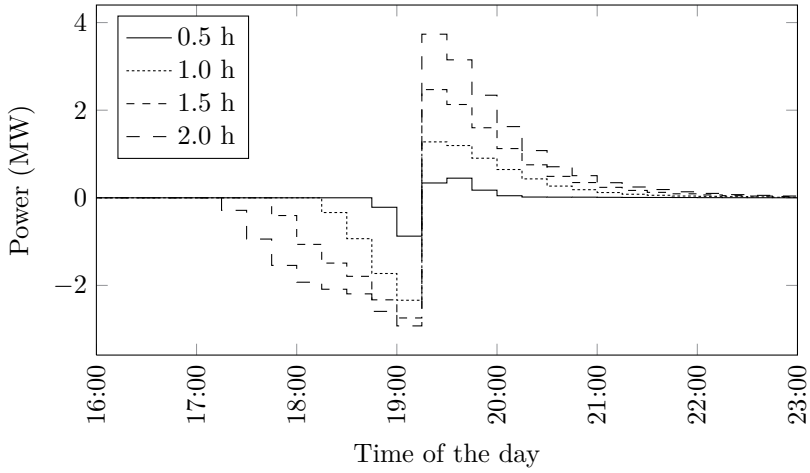



Figure 8.5: The impact of using tumble dryers' flexibility

Table 8.7: Peak before and after delay of dishwashers

| delay | participating [#] | power before [kW] | power after [kW] |
|---|---|---|---|
| 30' | 1460 | -1519 | 878 |
| 60' | 2558 | -1945 | 1509 |
| 90' | 3577 | -3227 | 2688 |
| 120' | 4008 | -3238 | 3356 |

The number of tumble dryers taking part in 30, 60, 90 and 120 minute delays are shown in Table 8.6 together with the impact before and after the delay.

### Dishwashers

The number of simulated dishwashers is 68 717, reflecting the lower ownership rates of dishwashers compared to washing machines and tumble dryers. Figure 8.6 shows the effect of delaying dishwashers. Dishwashers operate their heating resistors two times per cycle, explaining the two peaks and the negative low in between, power that would have been demanded from cycles that are delayed. This also explains why the power is relatively low when the dishwashers are allowed to start again and that the second peak is higher compared to the first one.

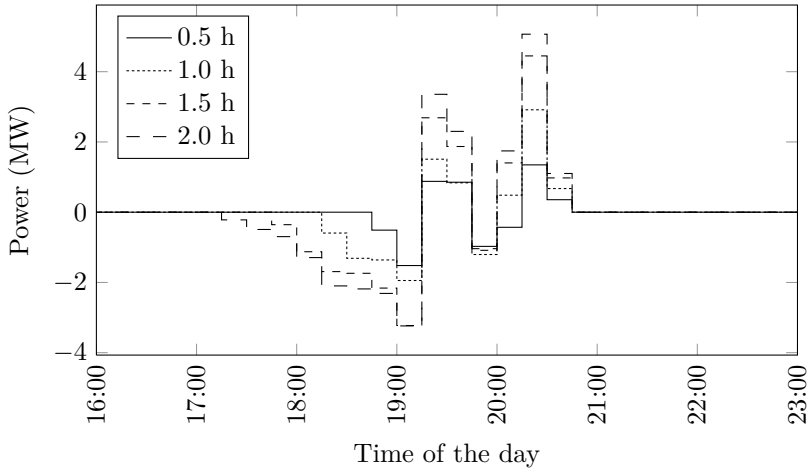



Figure 8.6: The impact of using dishwashers' flexibility

The number of dishwashers taking part in 30, 60, 90 and 120 minute delays are shown in Table 8.7 together with the impact before and after the delay.

### 8.2.2 Joint effect

The joint electricity demand of washing machines, tumble dryers and dishwashers is shown in Figure 8.7. The effect of delaying the starts on the joint demand is visible. Not allowing starts reduces the electricity demand drastically after half an hour. However, it takes long before the electricity demand of the appliances is gone.

At the moment the appliances are allowed to start again, power demand goes up. For a delay of one hour, the total power demand is 53.3 % higher. A longer delay (two hours) results in an increase by 142.5 %. One hour after the first peak, the power drops again. Dishwashers are responsible for this drop. The second peak is also caused by dishwashers.

A more detailed view on the effect of using the flexibility is depicted in Figure 8.8. For a delay of half an hour, the peak demand is relatively low because of the cycles that would have operated. The power demand before 19h15 is negative (Table 8.8). After 19h15 the power rises and remains about the same the next fifteen minutes. Half an hour after allowing starts, the power drops to slightly negative (412 kW), most dishwashers only require base load power after half

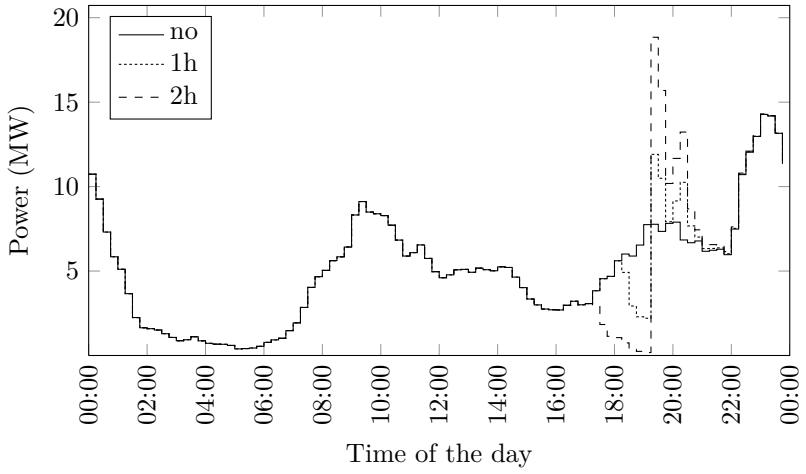



Figure 8.7: The effect of using flexibility on the total demand of wet appliances

Table 8.8: Peak before and after delay of wet appliances

| delay | participating [#] | power before [kW] | power after [kW] |
|---|---|---|---|
| 30’ | 3999 | -3977 | 1723 |
| 60’ | 7158 | -5543 | 4141 |
| 90’ | 10166 | -7311 | 7542 |
| 120’ | 13070 | -7573 | 11073 |

an hour. The base load combined with no power demand of cycles that would have operated explain the negative power.

A delay of one hour resulted in a higher negative just before and a more positive power just after 19h15 (Table 8.8). The double peaks are visible as well. The negative power is restricted to the avoided electricity demand, hence the same magnitude of negative power for one and a half hour and two hours (Table 8.8).

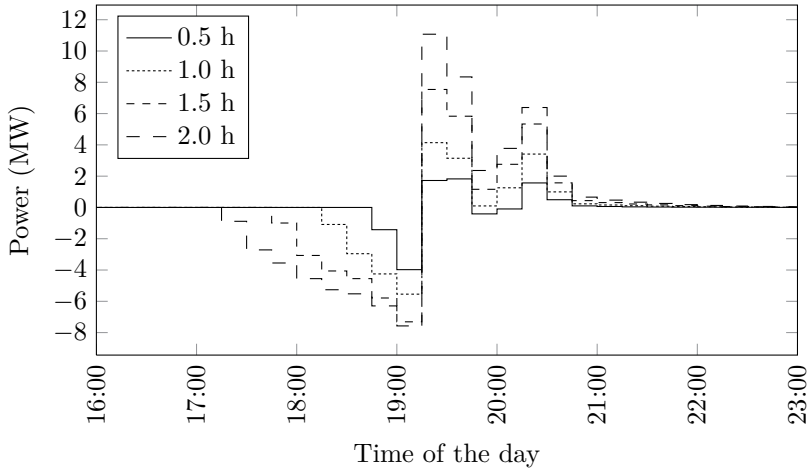



Figure 8.8: The impact of using wet appliances' flexibility

## 8.3 Conclusions

The impact of wet appliances on the total electricity demand is limited and ranges from 28 W (average business-consumers) to 87 W (relatively large day-consumers) on average. The wet appliances' electricity demand of the various customer groups is combined with the attitude towards active demand participation of the groups to estimate the potential. Only the most positive attitude is assumed to be willing to participate. The potential is found to be ranging from 11 W (large day-consumers) up to 44 W (relatively large day-consumers) on average.

The potential for active demand, i.e. flexibility, in Belgium is estimated by combining the probability of each customer group with the potential per group and multiplying this with the number of households. 29 % of the households are considered to be participating in active demand based on the most positive attitude.

The average potential for flexibility is expected to be 92 MW, with peaks up to 353 MW and is not negligible compared to the power reserves. However, it seems likely that the flexibility will only be used as a last resort for balancing because of the reserves' requirements regarding availability and response time.

The effect of using flexibility of wet appliances is tested by simulating 100 000 households for a Tuesday. The appliances are not allowed to start until 19h15, the disallowance varies from half an hour up to two hours with steps of a half

hour. The effect in terms of power reduction when the devices are delayed and the effect when the appliances are allowed to start again is described.

The peak demand when the appliances are allowed to start again is influenced by the delayed cycles, resulting in a relative lower peak for a small (fifteen minute) delay. The lower peak is encountered for every wet appliance. The power demand of dishwashers has an extra effect. Because of the two heating cycles during a dishwasher cycle, there is a peak first, followed by a drop in power, to end with a peak.

Longer delays involve more appliances. The negative power before starts are allowed again is however limited to the power demanded by the appliances which would have been started. The positive power on the other hand, scales up with an increasing delay and the increasing number of appliances.

The potential for active demand and the results of the simulations are based on multiple assumptions. The dimensions to cluster upon are assumed to be correct. The cluster membership is relaxed to spread the data over multiple customer groups, introducing errors. The customer group membership weight is used to scale up measurements of appliances. The number of appliances is limited, impacting the estimations based on measurements. Models are created to represent appliances and again assumptions are made. The results presented in this chapter are thus not exact, but the magnitude of the numbers is correct and give insights about the possibilities for active demand.

# Chapter 9

# Toolset

The Linear project, the major data source for the thesis, gathers data from different sources. Making data mining possible for multiple partners requires a specific software infrastructure, especially because of the confidentiality requirements of the parties providing data providing [151].

The requirements are listed in Section 9.1, the major ones being are related to security and functionality. Thereafter, an overview of the infrastructure is given (Section 9.2). How security and functionality are tackled, is described in Section 9.3. Some aspects of the work recently got improved by Strobbe et al. [14].

## 9.1 Requirements

The data in the Linear project is provided by various partners: two distribution system operators, two industrial partners, two research organisations and one (sub)metering company. Amongst the data are residential data of (Section 2.3), measurement data of wind farms and photovoltaic installations. The survey data (Section 2.4) are coupled with customers of the distribution system operators. The measurements and other data are gathered for the project. However, initially, multiple project partners provided measurement equipment. The partners aren't allowed to access each others data.

The industrial partners' main interest is the protection of their own data. Only the research institutions with a non-disclosure agreement are allowed to access the data or parts of the data. The researchers want availability and functionality.

### 9.1.1 Security

The security properties that need to be assessed are confidentiality, integrity, availability and accountability [122]. All are related to the asset to be protected: the data. Industrial partners and research institutions have a different view on the protection. The industrial partners are interested in confidentiality and accountability while the research institutions are focused on availability and integrity.

Confidentiality makes sure that only authorised people or software are allowed to access the content of the data. Certain parts of the source code of the software for the infrastructure need to be made inaccessible as well. The software code might provide information about the data structure to a competitor. Accountability is keeping track of who is responsible for which action. Holding someone accountable for leaking data is thus possible if the data is not kept confidential.

Integrity means that the data remains unaltered. Altering the data might result in invalid research conclusions. Availability refers to the time the data is accessible and the response time of the system.

### 9.1.2 Functionality

The functionality of the software in the infrastructure is important for the researchers. The infrastructure is used for data analysis, which means that it has to meet the requirements for 'knowledge discovery and data mining' (KDD) (Section 3.1).

Online analytical mining (OLAM) combines online analytical processing (OLAP) with data mining. The online analytical processing is used for the data transformation before data mining itself takes place. It should be possible to query the data with online analytical processing automatically.

The researchers should be able to adjust the data analysis tool, allowing them to implement their own data transformations and their own output format. The source code needs to be shared amongst researchers to improve productivity, i.e. no double implementations.

## 9.2 Overview

An overview of the infrastructure for data analysis is presented in Figure 9.1. The knowledge discovery and data mining process is divided in three stages:

- data cleaning and preprocessing,
- data storage and
- data transformation, analysis and interpretation.

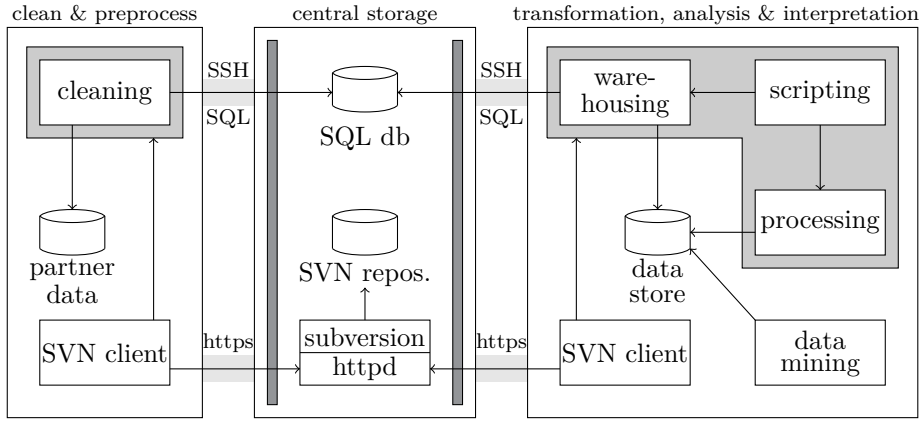



Figure 9.1: Data analysis infrastructure

Data cleaning and preprocessing is done locally at the site of an industrial partner. Each industrial partner is thus responsible for data cleaning and converting data into the project format (preprocessing). The data is placed directly into the central database. The distribution system operators provide 3 GiB of measurements. Renewable decentralized generation accounts for an additional 2 GiB of measurements. Linear dataset 1 is 3 GiB large. The raw data is formatted in comma separated value files.

All data is stored centrally, including the source code of the software. The database has restrictions: only authorised people are allowed to access predefined parts of the data. The same goes for the source code: not all source code is available to everyone.

The functionality for researchers is situated at the right hand side in Figure 9.1. A direct link with the central server ensures that they have access to the data and the source code repository. The software tool allows for data transformation in

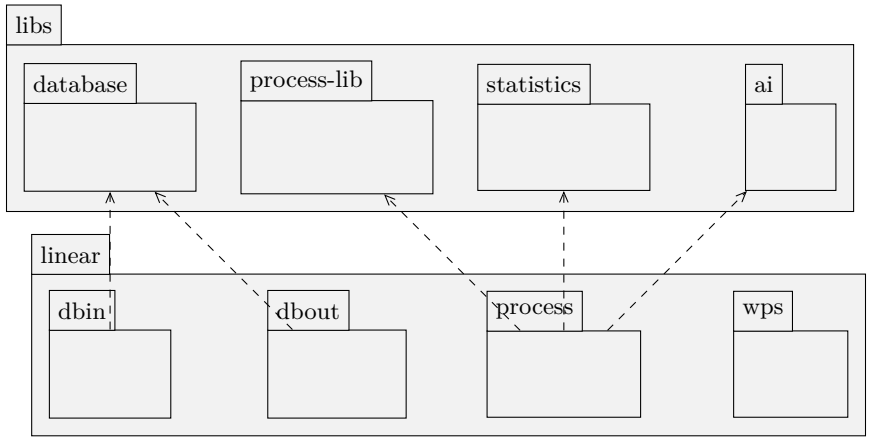



Figure 9.2: Software package overview

the form of query generation (warehousing) and numerical methods (processing). The results of the data transformation are stored locally. The processing functionality is also able to convert the data into the format of data mining tools. These tools are not developed in-house, but are the preferred data mining tools of the researchers.

The in-house developed software tool supporting cleaning, warehousing, scripting and processing is marked in grey (Figure 9.1). The package diagram of the software is shown in Figure 9.2. Two large packages can be distinguished: 'libs' and 'linear'. The 'libs' package contains all generic functionalities. The 'linear' package inherits from the 'libs' packages and implements the functionalities specific to the project. The generic package makes it possible to reuse the software in other projects, without giving away the specific implementations of the Linear project.

## 9.3 Implementation

Security requirements are implemented in the data analysis' infrastructure, functionalities on the other hand are mainly implemented in the software tool.

### 9.3.1 Security

The repository of Yskout et al. [126] is used to implement the security patterns in the infrastructure. The considered patterns are firewall, secure pipe, full view with errors and storing data at the client.

#### Firewall

A firewall protects the internal network from the exterior [152]. The pattern works at system level. Incoming and outgoing connections are restricted based on authorisations [126]: only connections from authorised IP-addresses are allowed.

The firewall is indicated in dark grey in Figure 9.1. Only IP-addresses from the research institutions and the industrial partners are regarded as authorised. Some institutions use dynamic IP-addresses. For them, IP-address ranges are defined.

The firewall also blocks ports: only secure shell (SSH, see secure pipe) and hypertext transfer protocol secure (HTTPS, see secure pipe) connections are allowed. Those are the only protocols needed in the infrastructure (Figure 9.1).

Network address blacklist [120] is a pattern related to the firewall pattern, but it operates in the opposite direction. With the firewall pattern, connections are restricted by default, while network address blacklist requires adding IP-addresses once suspicious behaviour is detected.

#### Secure pipe

Secure pipe operates at system level. Transport Layer Security (TLS) and Secure Sockets Layers (SSL) are considered to be secure pipes [126, 124]. A secure shell tunnel (SSH) [153] is not started from the application, but also creates a secure connection. Secure pipes protect against eavesdropping (confidentiality) and message tampering (integrity).

The SQL-queries (Structured Query Language) are tunnelled through a secure shell tunnel (Figure 9.1): the arrow represents the queries and the light grey pipes indicates the secure tunnel.

The use of SSH has extra advantages: the SQL-server only needs to listen to *localhost* (the server itself), increasing security. The SQL-server on the central server believes the requests come from the central server because of the SSH-tunnelling. The disadvantage of using secure shell is the requirement of an

account for each user. The users can thus place files on the server and execute code. To prevent users from trying to get administrative privileges, the root account must be disabled.

Asymmetric host keys are used in the implementation. The client stores the public RSA-key (Rivest, Shamir and Adleman) of the central server locally. The public key of the server is checked by the client for each connection. The client is notified if the key is invalid, which prevents man-in-the-middle attacks.

The version control system for the infrastructure is Apache Subversion. Apache Subversion operates centralised: there is only one repository, which makes it easier to track the source code. The version control system works together with the Apache web-server, allowing the use of Web Distributed Authoring and Versioning (WebDAV), operating with the hypertext transfer protocol (HTTP). Transport Layer Security (TLS) and Secure Sockets Layer (SSL) provide security for HTTP indicated by an arrow in Figure 9.1, while TLS/SSL are marked in light grey.

Hypertext transfer protocol secure (HTTPS) consists of TLS/SSL as a layer below HTTP. The communication cannot be eavesdropped nor tempered in this way. Man-in-the-middle attacks are prevented by working with a public key certificate of the central server, issued by a certification authority.

### Full view with errors

Full view with errors is a security pattern at application design: user are allowed a global view, but the system generates error messages once the user tries to perform an illegal action [124, 126]. The pattern helps to increase confidentiality. The counterpart of this pattern is limited view, where the user is only able to see the operations he or she is authorised to do.

The pattern, full view with errors, is used in the database. Each researcher has been allowed access to the relevant part of the data. The SQL-server generates an error message when the researcher tries to access data without have the permission. Researchers are only allowed to read out data, not to write data into the database.

The industrial partners only have access to their own data. However, because competing companies have to place data in the same tables, someone is appointed to place the data in the database. A better way to tackle this problem is by using web services, as implemented in the new version of the central server [14].

#### Storing data at the client

Storing data at the client increases its availability and reduces the server load. However, in general, it is preferred to store data at the server side as data from clients cannot always be trusted [120]. The client data storage pattern requires that the data is encrypted at the client side, but this would prevent researchers from doing their work.

The distinction in database access rights between industrial partners and researchers prevents data tampering: industrial partners are only allowed to change their own data and researchers have read-only access. Researchers are unable to alter data at the server, but they are free to work on the data locally. The availability increases: if the server goes down, the researchers are still able to work on the local data. The data is transformed (aggregated) before it is stored locally, as explained later. The data transformation increases confidentiality, as the original data is not stored.

The source code for the software tool is also copied locally. Researchers have write-access on the source code versioning system. The system keeps track of the committed changes to the source code. Changes can therefore be tracked back to a person, ensuring accountability.

Local data storage is indicated by 'partner data' and 'data store' in Figure 9.1. The source code is marked in grey in Figure 9.1 and is also copied locally. Industrial partners are able to keep specific cleaning routines local, i.e. not synchronising with the subversion repository. Competitors are thus unable to derive the internal data structure from the source code.

### 9.3.2 Functionality

The focus of the functionality is on the 'knowledge discovery and data mining'-process and automation. The functionality is mainly implemented in the software tool and written in Python. Python has good database support, a mathematical environment and is a scripting language which helps to automate.

#### Cleaning

The cleaning block in Figure 9.1 represents data cleaning and pre-processing in the knowledge discovery and data mining process. The data selection is done beforehand. Consensus between the industrial partners and the research institutions determined the data needed to be placed in the central server.

The ‘database’ package in Figure 9.2 provides generic functions to load data into a database. The ‘dbin’ package contains the database scheme of the ‘Linear’-project, together with routines to handle missing data and different time steps in the data. The data providing parties implement their own data cleaning and pre-processing routines, often specific to the data provider. They can be kept offline or can be submitted to the subversion repository.

### Warehousing

Data warehouses are databases used for analysis and reporting. Warehouses aggregate data by manipulating the dimensions of the data or by combining multiple sources of data. The same functionality is implemented in the ‘warehousing’-package (Figure 9.2), but data is stored locally (Figure 9.1) at the side of the researcher. Because of the local storage, the implementation is called warehousing and not data warehouse.

The warehousing module works as online analytical process (OLAP). Both relational and multidimensional OLAP are supported. Queries are used to process relational data in relational databases [154, 155]. In the multidimensional case, the queries are used to reduce the dimensionality of the data. Data is stored in special data structures (arrays and matrices) in multidimensional databases [154, 155]. Here, the multidimensional database is the local data store where comma separated value files are used to store arrays and matrices. Large parts of the multidimensional processing are also done by the processing module.

The purpose of the warehousing module is generating queries and storing the results in the local data store. The queries are mainly used to aggregate data. Structured Query Language (SQL) supports multiple aggregation functions: minimum, average, maximum, standard deviation, variance, etc. Date and time are stored in an standardised way. Various functions are available to work with data and time. An example of an automatically generated SQL-query is presented in Figure 9.3: the load curve of the average day of each month during the defined period is requested from all metering data.

The aggregation and the time handling by SQL was the reason to choose SQL. NoSQL databases also support SQL-like languages (e.g. Hive [156], Pig Latin [157] and HadoopDB [158]), but their focus larger datasets. The project has insufficient amount of data, i.e. terrabytes, to justify NoSQL technologies.

```
SELECT
  AVG(power) as p,
  month(ts) as mont, hour(ts) as hou, minute(ts) as minut
FROM
  metering
WHERE
  ts BETWEEN %s AND %s
GROUP BY
  mont, hou, minut
ORDER BY
  mont, hou, minut
```

Figure 9.3: An SQL-query to retrieve the load curve of the average day of each month for all metered profiles in the specified period

### Processing

The processing module handles the multidimensional online analytical processing that cannot be done by the warehousing library. The warehousing library is limited to relational processing and data aggregation in the time domain. The processing library includes functionality for statistics, algebraic transformations, plotting and extra date and time transformations. Figure 9.1 shows that the processing module is able to access the data store. Data is read from the local data store, processed and then stored back. The use of a the data store makes it possible to iterate over warehousing, processing, mining and processing again (visualisation).

From a package point of view, the processing module at library level is composed of a 'process', an 'ai' and a 'statistics' package (Figure 9.2). The 'process' package at project level implements data transformations and visualisations based on the various libraries. The routines and classes are implemented in the SciPy (Scientific Python) and NumPy (Numerical Python).

### Scripting

Designing and implementing a graphical user interface that supports automation is time consuming and complex. A scripting environment is more flexible as it gives the researchers a way to automate data retrieval and transformation. The transformations can easily be adapted to new insights. A scripting environment is thus more flexible. Another plus is that the scripts are exchangeable. The disadvantage for the researchers is having to know the scripting functions.

An example of a script is shown in Figure 9.4. Answers to a survey questions

```
hQuest = QuestionRange1()
hQuest.select = ["H_ID"]
hQuest.question1_1 = [1,2]
ids = hQuest.execute()

duration = {"from": datetime(2008,1,1),
  "to": datetime(2009,1,1) - timedelta(seconds=1)}
intervals = ["quarter","weekday","hour","minute"]

eMeasure = Measurements()
eMeasure.selection = ["E"]
eMeasure.duration = duration
eMeasure.aggregator = "avg"
eMeasure.intervals = intervals
eMeasure.h_ids = ids
eMeasure.getdata("output/measurements.csv")

readin = CSVReader("output/measurements.csv")
p = TimeSequencePlot(readin)
p.format = "png"
p.outputdir = "images/"
p.output = map(str, ids)
p.duration = duration
p.newgraphon = "weekday"
p.intervals = intervals
p.execute()
```

Figure 9.4: Scripting example

are the basis to select households. The electricity consumption measurements of 2008 from the selected households are converted into load curves. The load curves represent the average days of the week for the four quarters of the year. The result is stored in the data store. The saved load curves are loaded again and plotted as portable network graphics (PNGs) in the data store.

## 9.4 Conclusions

Allowing multiple researchers to work jointly on a project's dataset requires an a specific software infrastructure. The requirements for the infrastructure are related to security and functionality.

The data providing partners from industry are mainly concerned about confidentiality and accountability, while researchers value availability and data integrity. The security requirements are implemented with the help of security patterns, i.e. solutions to recurring information security problems [120].

The implemented patterns are ‘firewall’, ‘secure pipe’, ‘full view with errors’ and ‘storing data at the client’. The firewall protects the server from the outside: only a limited range of addresses is authorised to connect through a limited range of ports. The secure pipe ensures that the communication between the researcher or the data providing partner and the central server cannot be eavesdropped or tampered. Full view with errors gives a full overview of the data and the functionality, but generates errors when someone tries to access a function or data without having the authorisation. Storing data at the client is a modification of client data store. No encryption is applied, but the use of aggregated data improves confidentiality. Storing data at the client improves availability.

From a functionality point of view, the ‘knowledge discovery and data mining’-process needed to be supported by the infrastructure and the software tool. The source code of the software tool is shared on a central repository to make the exchange and the development of the framework easier. The software tool facilitates the ‘knowledge discovery and data mining’-process.

Data cleaning and preprocessing are done locally at the site of the data providing partners. The cleaned data is stored in a central database. Data transformation is performed by a warehousing and a processing module. The warehousing module generates queries to retrieve aggregated data from the database, while the processing module adds extra functionality to manipulate the data. The retrieved data is stored locally on the computer of the researcher. The preferred data mining tool of the researcher is able to read in the data and to store the data back after performing the analyses. Visualising the results is possible through functions in the processing module. No graphical user interface is designed, instead a scripting environment is used, being is more flexible: it makes automation easier and research scripts exchangeable. The cost of working with a scripting environment is the need to know the classes and the functions of the libraries.

# Chapter 10

# Conclusions and future work

The dissertation deals with description, selection and modelling of residential electricity demand, both at the connection point and of wet appliances. An infrastructure to perform data analyses has been designed and possible privacy issues are addressed. The conclusions of each chapter are compiled in this chapter, the contributions of this thesis are listed and suggestions are given for future research.

## 10.1 Conclusions

Simulations of smart grid cases, such as active demand response and impact of electrical vehicles on the grid, require representative data to draw conclusions. Residential data for smart grids consists of measurements and household information. Residential data is well protected by privacy laws: it is not always possible to obtain data and aggregation is often used to remove privacy sensitive information. Aggregation removes detailed information about the data, such as peaks, making it impossible to execute detailed simulations.

The first way to deliver data is by selection. To execute random sampling, the sampling frame (the set to sample from) needs to be representative to the whole population. However, due to non-response in surveys, the sampling frame is often biased. A technique to select customers or data is required.

To fix the bias, both information about demographic properties as well as electricity consumption of the population are required. The annual electricity consumption in Belgium is best represented by a skewed (Weibull) distribution

with a mean of 4.9 MWh, a mode of 2.8 MWh and a median of 3.6 MWh, the latter being the value the regulator uses for the average consumer. The demographic properties describing electricity demand best are age of the respondent, number of inhabitants, number of inhabitants being at home during the evening, surface area for business, owning a freezer or not, owning a dishwasher or not and housing type. The relations between demographic properties and electricity demand are found by a decision tree, performing better than an artificial neural network and a support vector machine.

Data selection is done by combining quota sampling with optimisation: automated quota sampling. Before selection, the data is made anonymous by removing direct links, i.e. name, address, to the customer. The quota domains are the found demographic parameters and the annual electricity demand. The optimisation algorithm ensures that uota requirements are met, mathematical programming performed better (i.e. faster, no permutations) in the task compared to constraint programming. The downside of quota sampling is the inability to examine the properties of the sampling estimators.

The second way for data delivery is by generation. Data selection might still be hampered by privacy problems and lack of data, data generation on the other hand generalises the data in models, removing privacy sensitive information and allowing for generating 'unlimited' data. To control output of data generation, e.g. customers with a low annual electricity demand behave differently from those with a high demand, groups of customers need to be defined.

A clustering algorithm (Expectation Maximisation) is applied to a set of customers (on average) representative of Flanders/Belgium. The clustering is based on timing and amount of electricity consumption, expressed by load curves. Ten clusters are found and named after their properties (timing, magnitude and having a business), each having a representative, consisting of a load curve representing the average demand of the group and the standard deviations over the load curve. The shape of the load curves resemble the synthetic load profiles of the regulator, who used the same data. The cluster membership can be relaxed, making it possible to spread customer data over various clusters, scaling up the data. The downside is the change in the distribution of the data and the changes of the probabilities of the clusters.

Appliance measurements of only a small set of customers are available. The relaxed cluster membership is used to spread the data of those customers over the various customer groups, creating a load curve of the appliance usage in the cluster. The shape of the load curves of washing machines and tumble dryers is comparable to the load curves described in the literature. However, mainly the load curves of dishwashers are influenced by the tariff structures in Belgium and hence differ from the literature. The scaled up small set of measurements

is thus able to represent the electricity demand of wet appliances.

Markov chains provide ways to model autocorrelated data such as residential load profiles: measurement data are split up in states and probabilities between those states are calculated. The state boundaries are based on the electrical power distributions of the relaxed clusters, the probabilities of the transitions on frequency of occurring and the relaxed cluster membership. Behaviour Markov chains keep the behaviour consistent, variation Markov chains add variation to the daily behaviour.

Load profiles are generated by taking single step samples from the Markov chains. Load curves created from the generated profiles have, compared to the original cluster load curves, a similar shape, ensured by the behaviour Markov chain, and a lower average power, because of the Markov property. Autocorrelation could be reproduced, because of the combination of behaviour and variation to behaviour. In simulations using the generated profiles, electricity demand will be lower and peaks will be somewhat lower and less frequent. However, the overall electricity demand will be similar to the original data for most clusters. The relaxed model is corrected with the original cluster probabilities and Markov models are trained with those corrected cluster memberships: electricity demand is closer to the original and peaks are higher compared to the relaxed.

Wet appliances are modelled by two properties: starts and settings. Starts are aggregated into start curves per cluster. Probability distributions model appliances' settings per cluster. In both cases, the relaxed model is applied to spread the limited number of appliances over the various clusters.

Load cycles generated from the clusters' start curves and distributions are aggregated into load curves and compared to the clusters' load curves of the measured load cycles. The average power of the generated is lower than the average power of the original load profiles of the wet appliances. The shape of the generated is comparable to the original load curve, but with lower peaks. Again, in simulations, appliance electricity demand will be lower and peaks will be somewhat lower and less frequent. The appliance overall electricity demand will be similar to the original appliance data. No correction is applied: the relaxed cluster membership implies that the cluster models are correct (the assumption made by making the appliance models), while the corrected cluster memberships push the data closer to the original cluster data.

Most simulations with the selected or generated data described above are related to (active) demand response. To indicate what to expect, estimations about the impact of appliances on the total demand as well as the potential and the effect of using flexibility are explored. The average impact of wet appliances is limited and ranges from 28 W up to 87 W per household, depending on the cluster

(i.e. customer group). Combining the impact with the attitude towards active demand response results in the potential: ranging from 11 W up to 44 W on average per household, again depending on the cluster. In terms of the whole Belgian population (4.6 million households), the potential is expected to be 92 MW on average with peaks up to 353 MW. The numbers require 29 % of the households or 1.3 million households to participate. The magnitude is in the order of the power reserves, but doesn't fully meet the requirements for availability and response time.

The effect of using active demand is tested by simulating appliances of 100 000 participating households. Negative power can be created by delaying appliances: it will take at least fifteen minutes before the negative power is established, the corresponding power increase afterwards will be lower if the delay is smaller than one and a half hour. To create power peaks, long delays are required: longer delays involve more appliances, negative power is limited by the appliances that would have been started at that time, positive power scales with an increasing delay.

The possible privacy issues related to the detection and monitoring of appliances have been assessed as well: 'How much extra information, e.g. income, can be derived from monitoring systems?' Distribution system operators will monitor the electricity demand at the point of common coupling of houses with smart meters. Detecting appliances in those measurements is only possible if the resolution is high enough, for example 6 seconds. Appliances are hard to detect on a fifteen minute time scale, the resolution of the smart meters. Hence, distribution system operators will have difficulty in detecting appliances. Providers of home management systems, on the other hand, monitor the electricity demand of individual appliances as well as the electricity demand at the connection point on a fifteen minute scale. The cycle detection and settings estimation algorithms for washing machines, tumble dryers and dishwashers presented allow for the estimation of privacy sensitive information, such as income and family expansion.

A software infrastructure has been designed and implemented to facilitate the above analyses. Security patterns are used to ensure confidentiality, availability, integrity and accountability. The functionality is designed around the 'knowledge discovery and data mining'-process.

The main contributions can be summarised as follows:

- A description residential electricity demand is presented by the distribution of annual electricity demand and the related demographic parameters.

- Quota optimisation is proposed as a method to select customers from a set, biased due to non-response.
- Customers are grouped based on timing and magnitude of electricity demand. Cluster representatives can be used for fuzzy clustering, allowing to scale data up.
- Markov chains are used to model residential load profiles.
- Wet appliance usage is modelled by start curves and settings distributions.
- Possible privacy issues related to the detection of appliances are explained.
- A framework for data-analysis is implemented.

## 10.2 Future work

The thesis deals with the description, the selection and the modelling of residential electricity demand at the connection point of a house and of wet appliances. The models to select or generate electrical load profiles are intended to be be used in simulations and to gain insights in electricity demand in Belgium. The generation of load profiles of households can be coupled to a distribution net to, for example, estimate the impact of electrical vehicles on the grid.

The models of the wet appliances allow for the simulation of shifting wet appliances for a range of technical and/or economical objectives. The economical feasibility of implementing active demand with wet appliances is one of the most important questions.

The data-analyses infrastructure is written in a generic way and can be applied to other projects. Improvements of the infrastructure are already implemented and used in the field-test of the 'Linear'-project. The software for the analyses itself however, is still used.

Next to applications of the results, improvements can be researched as well. The clustering of electrical load profiles is described extensively in the literature: various clustering algorithms algorithms and data transformations are used. Load curves as data transformation and Expectation Maximisation clustering as algorithms are the basis for this work. However, other transformations and clustering techniques might generate better results.

The number of dimensions to cluster upon is chosen in order to include the influence of the seasons and the days of the week. The relaxation reduces the

dimensionality afterwards, but introduces an error. A clustering technique that is able to perform fuzzy clustering, where the weights are better able to spread the data over the clusters, without having to reduce the dimensionality afterwards is an interesting follow-up topic.

The description of electricity demand of wet appliances is based on a small set of measurements. More measurements will improve the description. The inclusion of other appliances should be considered as well. A better clustering algorithm, as mentioned above, could also improve the description of the electricity demand by appliances.

The detection and settings estimation of the wet appliances is done based on the manuals of the corresponding appliances. Measuring each appliance in detail will give more insights into the working principle of the appliances and will result a better detection and settings estimation. Better settings detections will allow for a better estimation of privacy related information and better models for the appliances.

The security of the data-analyses infrastructure makes use of external programs. Security can also be built in into the software by defining interfaces to retrieve data from the central server. The software has to be placed on the central server to support this and requires that the software is mature.

# Bibliography


[1] R. Belmans, J. Driesen, G. Deconinck, G. Vekemans, G.-J. Schaeffer, E. Peeters, P. De Meester, J. Poortmans, D. Vanderzande, J. Declercq, G. Palmers, L. Dewilde, Y. Vanlinthout, L. Vandenbulcke, A. Vermeylen, and F. Couttenier, "Naar wereldleiderschap in hernieuwbare energietechnologie en elektrische infrastructuur, Strategische voorstellen aan de Vlaamse overheid," 2009.

[2] B. Dupont, P. Vingerhoets, P. Tant, K. Vanthournout, W. Cardinaels, T. De Rybel, E. Peeters, and R. Belmans, "Linear breakthrough project: Large-scale implementation of smart grid technologies in distribution grids," in *Innovative Smart Grid Technologies (ISGT Europe), 2012 3rd IEEE PES International Conference and Exhibition on*, Berlin, Germany, 2012.

[3] J. Rowley, "The wisdom hierarchy: representations of the dikw hierarchy," *Journal of Information Science*, vol. 33, no. 2, pp. 163 – 180, 2007.

[4] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, 2005.

[5] *Ediel Model voor de geliberaliseerde energiesector in België*, Synergrid, Math-X N.V. and UMIX-Arevi, 2011. [Online]. Available: http://www.atrias.be/NL/Publications_Atrias/02%20Handbooks/Handbook%20SLP%202012.pdf

[6] VREG, "Beslissing van de Vlaamse Regulator van de Elektriciteits- en Gasmarkt met betrekking tot de goedkeuring van de categorieën van de distributienetgebruikers zonder registratie van het gemeten verbruiksprofiel en de overeenstemmende synthetische lastprofielen voor het jaar 2008, zoals bedoeld in de artikelen 3.3.3. en 3.3.4 van de meetcode van het technisch reglement distributie elektriciteit." Nov. 2012. [Online]. Available: http://www.vreg.be/verbruiksprofielen-0

[7] W. Labeeuw, C. Mol, and S. Claessens, "D1.1: Rapport analyse bestaande meetgegevens en studies," 2010, linear-project internal document.

[8] *NPS-exportcatalogus*, index,is, 2006.

[9] VREG, "Persmededeling van de Vlaamse Reguleringsinstantie voor de Elektriciteits- en Gasmarkt van 4 juni 2009 met betrekking tot de lancering van de VREG-zonnepanelen website, de achterstand bij het behandelen van aanvragen van PV-eigenaars en de 20.000ste PV-installatie in Vlaanderen," Jun. 2009. [Online]. Available: http://www.vreg.be/sites/default/files/persmededelingen/pers-2009-8.pdf

[10] ——, "Evolutie van het aantal zonnepanelen en hun vermogen," Okt. 2013. [Online]. Available: http://www.vreg.be/sites/default/files/uploads/stand_van_zaken_op_1_oktober_2013.pdf

[11] J. Stragier, L. De Marez, and L. Hauttekeete, "D1.2.1.a: Rapport gebruikersstudie: Inzichten, attitude en gedrag met betrekking tot energy management," 2010, linear-project internal document.

[12] Directorate General for Statistics and Economic Information Belgium, "Structuur van de bevolking," 2011. [Online]. Available: http://statbel.fgov.be/

[13] A. Brooks, E. Lu, D. Reicher, C. Spirakis, and B. Weihl, "Demand dispatch," *IEEE Power and Energy Magazine*, vol. 8, no. 3, pp. 20 – 29, 2010.

[14] M. Strobbe, T. Verschueren, S. Melis, D. Verslype, K. Mets, F. De Turck, and C. Develder, "Design of a management infrastructure for smart grid pilot data processing and analysis," in *Proceedings of the IFIP/IEEE International Symposium on Integrated Network Management*, Ghent, Belgium, 2013.

[15] S. Iacovella, P. Tant, and G. Deconinck, "Lessons learnt from the linear large-scale energy monitoring field test," in *Proceedings of the IEEE Benelux Young Researchers Symposium In Electrical Power Engineering 2012*, Delft, The Netherlands, 2012.

[16] G. Piatetsky-Shapiro, "Knowledge discovery in real databases: A report on the IJCAI-89 workshop," *AI Magazine*, vol. 11, no. 5, pp. 68 – 70, 1991.

[17] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, pp. 37 – 54, 1996.

[18] ——, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27 – 34, 1996.

[19] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199 – 231, 2001.

[20] H. Blockeel, *Machine Learning and Inductive Inference*, 2nd ed. Acco, 2008.

[21] T. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill, 1997.

[22] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[23] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, pp. 68 – 80, 2012.

[24] K. A. J. Doherty, R. G. Adams, and N. Davey, "Unsupervised learning with normalised data and non-euclidean norms," *Applied Soft Computing*, vol. 7, pp. 203 – 210, 2007.

[25] W. Winstor, *Operations Research: Applications and Algorithms*, 3rd ed. Duxbury Press, 1994.

[26] I. Kovalenko, N. Kuznetsov, and V. Shurenkov, *Models of Random Processes, a handbook for Mathematicians and Engineers*. CRC Press, 1996.

[27] P. Levy and S. Lemeshow, *Sampling of Populations: Methods and Applications*, 2nd ed. Wiley, 1991.

[28] W. Trochim and P. Donnelly, *The Research Methods Knowledge Base*, 3rd ed. Atomic Dog Publishing, 2006.

[29] G. Kalton, *Introduction to Survey Sampling*. Sage, 1983.

[30] H. Jeeninga, M. Uyterlinde, and J. Uitzinger, "Energieverbruik van energiezuinige woningen," ECN, Tech. Rep. ECN-C-01-072, 2001.

[31] P. Boonekamp, "Improved methods to evaluate realised energy savings," Ph.D. dissertation, University of Utrecht, 2005.

[32] C. Vringer, "Analysis of the energy requirement for household consumption," Ph.D. dissertation, University of Utrecht, 2005.

[33] E. de Groot, M. Spiekman, and I. Opstelten, "361: Dutch research into user behaviour in relation to energy use of residences," in *PLEA 2008 - 25th Conference on Passive and Low Energy Architecture*, Dublin, 2008.

[34] I. Mansouri, M. Newborough, and P. Douglas, "Energy consumption in UK households: Impact of domestic electrical appliances," *Applied Energy*, vol. 54, no. 3, pp. 211 – 285, 1996.

[35] S. Firth, K. Lomas, A. Wright, and R. Wall, "Identifying trends in the use of domestic appliances from household electricity consumption measurements," *Energy and Buildings*, vol. 40, no. 5, pp. 926 – 936, 2008.

[36] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 933 – 940, 2006.

[37] G. Tsekouras, N. Hatziargyriou, and E. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1120 – 1128, 2007.

[38] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, "A new index and classification approach for load pattern analysis of large electricity customers," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 153 – 160, 2012.

[39] G. Coke and M. Tsao, "Random effects mixture models for clustering electrical load series," *Journal of Time Series Analysis*, vol. 31, no. 6, pp. 451–464, 2010.

[40] M. Sforna, "Data mining in a power company customer database," *Electric Power Systems Research*, vol. 55, no. 3, pp. 201 – 209, 2000.

[41] S. Verdu, M. Garcia, C. Senabre, A. Marin, and F. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1672 – 1682, 2006.

[42] S. Valero, M. Ortiz, C. Senabre, C. Alvarez, F. Franco, and A. Gabaldon, "Methods for customer and demand response policies selection in new electricity markets," *Generation, Transmission Distribution, IET*, vol. 1, no. 1, pp. 104 – 110, 2007.

[43] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer classification and load profiling method for distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, 2011.

[44] E. Carpaneto, G. Chicco, R. Napoli, and M. Scutariu, "Electricity customer classification using frequency–domain load pattern data," *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 1, pp. 13 – 20, 2006.

[45] M. Jin, H. Renmu, and D. Hill, "Load modeling by finding support vectors of load data from field measurements," *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 726–735, 2006.

[46] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 1232 – 1239, 2004.

[47] R. Stamminger, *Synergy Potential of Smart Domestic Appliances in Renewable Energy Systems*. Shaker, 2009.

[48] M. Presutto, R. Stamminger, R. Scialdoni, L. Cutaia, W. Mebane, and R. Esposito, "Preparatory studies for eco-design requirements of EuPs, LOT 14: Domestic washing machines and dishwashers task 3 - 5," 2007, (Tender TREN/D1/40-2005).

[49] European Commission, Directorate-General for Energy and Transport, "Green paper on energy efficiency, doing more with less," 2005.

[50] P. Berkholz, A. Brückner, A. Kruschwitz, and R. Stamminger, *Definition und Ermittlung verhaltensabhängiger Energieeinsparpotentiale beim Betrieb elektrischer Haushaltwaschmaschinen*. Shaker-Verlag, 2007.

[51] C. Cuijpers and B.-J. Koops, "Het wetsvoorstel "slimme meters": een privacytoets op basis van art. 8 evrm," Universiteit van Tilburg, Centrum voor Recht, Technologie en Samenleving, Tech. Rep., 2008.

[52] B.-J. Koops and C. Cuijpers, "Begluren en besturen door slimme energiemeters: Een ongerecht vaardigde inbreuk op onze privacy," *Privacy en Informatie*, vol. 6, no. 1, pp. 2 – 7, 2009.

[53] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1249 – 1260, 1992.

[54] L. Norford and S. Leeb, "Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms," *Energy and Buildings*, vol. 24, no. 1, pp. 51 – 64, 1996.

[55] M. Zeifman, "Disaggregation of home energy display data using probabilistic approach," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 1, pp. 23 – 31, 2012.

[56] M. Lisovich, D. Mulligan, and S. Wicker, "Inferring personal information from demand-response systems," *Security and Privacy, IEEE*, vol. 8, no. 1, pp. 11 – 20, 2010.

[57] J. Powers, B. Margossian, and B. Smith, "Using a rule-based algorithm to disaggregate end-use load profiles from premise-level data," *IEEE Computer Applications in Power*, vol. 4, no. 2, pp. 42 – 47, 1991.

[58] L. Farinaccio and R. Zmeureanu, "Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses," *Energy and Buildings*, vol. 30, no. 3, pp. 245 – 259, 1999.

[59] M. Marceau and R. Zmeureanu, "Nonintrusive load disaggregation computer program to estimate the energy consumption of major end uses in residential buildings," *Energy Conversion and Management*, vol. 41, no. 13, pp. 1389 – 1403, 2000.

[60] J. Kolter, S. Batra, and A. Ng, "Energy disaggregation via discriminative sparse coding," in *Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2011.

[61] S. Leeb, S. Shaw, and J. Kirtley, J.L., "Transient event detection in spectral envelope estimates for nonintrusive load monitoring," *IEEE Transactions on Power Delivery*, vol. 10, no. 3, pp. 1200 – 1210, 1995.

[62] W. Wichakool, A. T. Avestruz, R. Cox, and S. Leeb, "Modeling and estimating current harmonics of variable electronic loads," *IEEE Transactions on Power Electronics*, vol. 24, no. 12, pp. 2803 – 2811, 2009.

[63] D. Srinivasan, W. S. Ng, and A. Liew, "Neural-network-based signature recognition for harmonic source identification," *IEEE Transactions on Power Delivery*, vol. 21, no. 1, pp. 398 – 405, 2006.

[64] H. Y. Lam, G. S. K. Fung, and W. K. Lee, "A novel method to construct taxonomy electrical appliances based on load signaturesof," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 653 – 660, 2007.

[65] J. Liang, S. Ng, G. Kendall, and J. Cheng, "Load signature study - part i: Basic concept, structure, and methodology," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 551 – 560, 2010.

[66] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 76 – 84, 2011.

[67] T. Ueno, F. Sano, O. Saeki, and K. Tsuji, "Effectiveness of an energy-consumption information system on energy savings in residential houses based on monitored data," *Applied Energy*, vol. 83, no. 2, pp. 166 – 183, 2006.

[68] S. Park, H. Kim, H. Moon, J. Heo, and S. Yoon, "Concurrent simulation platform for energy-aware smart metering systems," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1918 – 1926, 2010.

[69] J. Han, C.-S. Choi, and I. Lee, "More efficient home energy management system based on zigbee communication and infrared remote controls," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 85 – 89, 2011.

[70] E. Quinn, "Privacy and the new energy infrastructure," Center for Energy and Environmental Security (CEES), Tech. Rep., 2008.

[71] E. McKenna, I. Richardson, and M. Thomson, "Smart meter data: Balancing consumer privacy concerns with legitimate applications," *Energy Policy*, vol. 41, pp. 807 – 814, 2012.

[72] A. Rial and G. Danezis, "Privacy-preserving smart metering," in *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, ser. WPES '11. Chicago, Illinois, USA: ACM, 2011, pp. 49 – 60.

[73] E. Steel, C. Locke, E. Cadman, and B. Freese, "How much is your personal data worth?" *Financial Times*, June 12 2013. [Online]. Available: http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html#axzz2XWNJ2mLX

[74] T. Hargreaves, M. Nye, and J. Burgess, "Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors," *Energy Policy*, vol. 38, no. 10, pp. 6111 – 6119, 2010.

[75] A. Faruqio, R. Hledik, and S. Sergici, "Piloting the smart grid," *The Electricity Journal*, vol. 22, no. 7, 2009.

[76] C. Bartusch, F. Wallin, M. Odlare, I. Vassileva, and L. Wester, "Introducing a demand-based electricity distribution tariff in the residential sector: Demand response and customer perception," *Energy Policy*, vol. 39, no. 9, pp. 5008 – 5025, 2011.

[77] H. Saele and O. Grande, "Demand response from household customers: Experiences from a pilot study in Norway," *IEEE Transactions on Smart Grid*, vol. 2, no. 1, pp. 102–109, march 2011.

[78] W. Labeeuw and G. Deconinck, "Residential electrical load model based on mixture model clustering and markov models," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 3, pp. 1561 – 1569, 2013.

[79] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi, "A bottom-up approach to residential load modeling," *IEEE Transactions on Power Systems*, vol. 9, no. 2, pp. 957 – 964, 1994.

[80] M. Stokes, "Removing barriers to embedded generation: a fine-grained load model to support low voltage network performance analysis," Ph.D. dissertation, Institute of Energy and Sustainable Development, De Montfort University, Leicester, 2005.

[81] R. Yao and K. Steemers, "A method of formulating energy load profile for domestic buildings in the uk," *Energy and Buildings*, vol. 37, no. 6, pp. 663 – 671, 2005.

[82] J. Widén, M. Lundh, I. Vassileva, E. Dahlquist, K. Ellegård, and E. Wäckelgård, "Constructing load profiles for household electricity and hot water from time-use data-modelling approach and validation," *Energy and Buildings*, vol. 41, no. 7, pp. 753 – 768, 2009.

[83] I. Richardson, M. Thomson, and D. Infield, "A high-resolution domestic building occupancy model for energy demand simulations," *Energy and Buildings*, vol. 40, no. 8, pp. 1560 – 1566, 2008.

[84] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy and Buildings*, vol. 42, no. 10, pp. 1878 – 1887, 2010.

[85] F. McLoughlin, A. Duffy, and M. Conlon, "The generation of domestic electricity load profiles through markov chain modelling," in *3rd International Scientific Conference on Energy and Climate Change*, Athens, Greece, 2010, pp. 18 – 27.

[86] U.S. Department of Energy, "Benefits of demand response in electricity markets and recommendations for achieving them, a report to the United States Congress pursuent to Section 1252 of the Energy Policy Act of 2005," February 2006.

[87] L. Hancher, X. He, I. Azevedo, N. Keyaerts, L. Meeus, and J.-M. Glachant, "THINK topic 11: 'shift, not drift: Towards active demand response and beyond'," European University Institute, Tech. Rep., 2013.

[88] D. Kirschen, "Demand-side view of electricity markets," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 520 – 527, 2003.

[89] G. Strbac, "Demand side management: Benefits and challenges," *Energy Policy*, vol. 36, no. 12, pp. 4419 – 4426, 2008.

[90] J. Torriti, M. G. Hassan, and M. Leach, "Demand response experience in Europe: Policies, programmes and implementation," *Energy*, vol. 35, no. 4, pp. 1575 – 1583, 2010.

[91] J. Kim and A. Shcherbakova, "Common failures of demand response," *Energy*, vol. 36, no. 2, pp. 873 – 880, 2011.

[92] Y. He, B. Wang, J. Wang, W. Xiong, and T. Xia, "Residential demand response behavior analysis based on monte carlo simulation: The case of Yinchuan in China," *Energy*, vol. 47, pp. 230 – 236, 2012.

[93] K. Hamilton and N. Gulhar, "Taking demand response to the next level," *IEEE Power and Energy Magazine*, vol. 8, no. 3, pp. 60–65, May-Jun. 2010.

[94] G. C. Heffner, "Configuring load as a resource for competitive electricity markets–review of demand response programs in the U.S. and around the world," in *Proceedings of the 14th Annual Conference of the Electric Power Supply Industry (CEPSI), Fukuoka*, 2002.

[95] C. Álvarez Bel, M. Alcázar Ortega, G. Escrivá Escrivá, and A. Gabaldón Marín, "Technical and economical tools to assess customer demand response in the commercial sector," *Energy Conversion and Management*, vol. 50, no. 10, pp. 2605 – 2612, 2009.

[96] I. Rowlands, D. Scott, and P. Parker, "Consumers and green electricity: profiling potential purchasers," *Business Strategy and the Environment*, vol. 12, no. 1, pp. 36 – 48, 2003.

[97] A. Faruqui, S. George, and J. Winfield, "Quantifying customer response to dynamic pricing," *The Electricity Journal*, vol. 18, no. 4, pp. 53 – 63, 2005.

[98] G. Pepermans, "The value of continuous power supply for flemish households," *Energy Policy*, vol. 39, no. 12, pp. 7853 – 7864, 2011.

[99] T. Ericson, "Short-term electricity demand response," Ph.D. dissertation, Norwegian University of Science and Technology, Department of Electrical Power Engineering, 2007.

[100] P. Finn, M. O'Connell, and C. Fitzpatrick, "Demand side management of a domestic dishwasher: Wind energy gains, financial savings and peak-time load reduction," *Applied Energy*, vol. 101, pp. 678 – 685, 2013.

[101] F. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319 – 340, 1989.

[102] F. Davis, R. Bagozzi, and P. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models," *Management Science*, vol. 35, no. 8, pp. 982 – 1003, 1989.

[103] N. Leemput, J. Van Roy, F. Geth, P. Tant, B. Claessens, and J. Driesen, "Comparative analysis of coordination strategies for electric vehicles," in *Innovative Smart Grid Technologies (ISGT Europe), 2011 2nd IEEE PES International Conference and Exhibition on*, 2011.

[104] S. Vandael, B. Claessens, M. Hommelberg, T. Holvoet, and G. Deconinck, "A scalable three-step approach for demand side management of plug-in hybrid vehicles," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 720 – 728, 2013.

[105] K. Mets, R. D'hulst, and C. Develder, "Comparison of intelligent charging algorithms for electric vehicles to reduce peak load and demand variability in a distribution grid," *Journal of Communications and Networks*, vol. 14, no. 6, pp. 672 – 681, 2012.

[106] I. Atzeni, L. Ordonez, G. Scutari, D. Palomar, and J. Fonollosa, "Noncooperative and cooperative optimization of distributed energy generation and storage in the demand-side of the smart grid," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2454 – 2472, 2013.

[107] T. Logenthiran, D. Srinivasan, and T. Z. Shun, "Demand side management in smart grid using heuristic optimization," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1244 – 1252, 2012.

[108] Z. Chen, L. Wu, and Y. Fu, "Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1822 – 1831, 2012.

[109] M. Pipattanasomporn, H. Feroze, and S. Rahman, "Multi-agent systems in a distributed smart grid: Design and implementation," in *Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES*, 2009.

[110] M. Pedrasa, T. Spooner, and I. MacGill, "Coordinated scheduling of residential distributed energy resources to optimize smart home energy services," *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 134 – 143, 2010.

[111] Y. Xu, W. Liu, and J. Gong, "Stable multi-agent-based load shedding algorithm for power systems," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2006 – 2014, 2011.

[112] J. K. Kok, C. J. Warmer, and I. G. Kamphuis, "Powermatcher: multiagent control in the electricity infrastructure," in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, ser. AAMAS '05. New York, NY, USA: ACM, 2005, pp. 75 – 82.

[113] K. De Craemer and G. Deconinck, "Balancing trade-offs in coordinated PHEV charging with continuous market-based control," in *Innovative Smart Grid Technologies (ISGT Europe), 2012 3rd IEEE PES International Conference and Exhibition on*, 2012.

[114] K. De Craemer, S. Vandael, B. Claessens, and G. Deconinck, "An event-driven dual coordination mechanism for demand side management of phevs," *IEEE Transactions on Smart Grid*, accepted, to be published.

[115] S. Rusitschka, K. Eger, and C. Gerdes, "Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain," in *First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct. 2010, pp. 483 – 488.

[116] F. Chong, "Multi-tenant data architecture," Microsoft, Tech. Rep., 2006. [Online]. Available: http://msdn.microsoft.com/en-us/library/aa479086.aspx

[117] A. Bâra, I. Lungu, M. Velicanu, and S. Oprea, "Intelligent systems for predicting and analyzing data in power grid companies," in *Information Society (i-Society), 2010 International Conference on*, Jun. 2010, pp. 266 – 271.

[118] J. Liu, X. Li, D. Liu, H. Liu, and P. Mao, "Study on data management of fundamental model in control center for smart grid operation," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 573 – 579, 2011.

[119] J. Chen, W. Li, A. Lau, J. Cao, and K. Wang, "Automated load curve data cleansing in power systems," *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 213 – 221, sept. 2010.

[120] D. Kienzle, M. Elder, D. Tyree, and J. Edwards-Hewitt, "Security patterns repository, version 1.0," 2006. [Online]. Available: http://www.scrypt.net/~celer/securitypatterns/repository.pdf

[121] T. Heyman, K. Yskout, R. Scandariato, and W. Joosen, "An analysis of the security patterns landscape," in *Software Engineering for Secure Systems, 2007. SESS '07: ICSE Workshops 2007. 3rd International Workshop on*, May 2007.

[122] N. Yoshioka, H. Washizaki, and K. Maruyama, "A survey on security patterns," *Progress in Informatics*, no. 5, pp. 35 – 47, 2008.

[123] A. Yautsiukhin, R. Scandariato, T. Heyman, F. Massacci, and W. Joosen, "Towards a quantitative assessment of security in software architectures," in *Nordic Workshop on Secure IT Systems (NordSec)*, Oct. 2008.

[124] J. W. Yoder and J. Barcalow, "Architectural patterns for enabling application security," in *Fourth Conference on Pattern Languages of Programs (PLoP 1997)*, Monticello, Illinois, Sept. 1997.

[125] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1995.

[126] K. Yskout, T. Heyman, R. Scandariato, and W. Joosen, "A system of security patterns," DistriNet Research Group, Katholieke Universiteit Leuven, Tech. Rep., Januari 2007.

[127] C. De Jonghe, B. Hobbs, and R. Belmans, "Optimal generation mix with short-term demand response and wind penetration," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 830 – 839, 2012.

[128] K. Dyke, N. Schofield, and M. Barnes, "The impact of transport electrification on electrical networks," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 12, pp. 3917 – 3926, 2010.

[129] K. Clement, E. Haesen, and J. Driesen, "The impact of vehicle-to-grid on the distribution grid," *Electric Power Systems Research Journal*, vol. 81, no. 1, pp. 185 – 192, 2011.

[130] H. Kanchev, D. Lu, F. Colas, V. Lazarov, and B. Francois, "Energy management and operational planning of a microgrid with a PV-based active generator for smart grid applications," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 10, pp. 4583 – 4592, 2011.

[131] J. Tant, F. Geth, D. Six, P. Tant, and J. Driesen, "Multiobjective battery storage to improve pv integration in residential distribution grids," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 1, pp. 182 – 191, 2013.

[132] VREG, "Welk type verbruiker bent u ?" 2013. [Online]. Available: http://www.vreg.be/welke-verbruiker-bent-u

[133] C. Versluys and B. De Wispelaere, private communication, SPE, 2011.

[134] N. Hastings and J. Peacock, *Statistical Distributions*. Wiley, 1975.

[135] G. Kanji, *100 Statistical Tests*. Sage, 1993.

[136] R. Bellman, *Adaptive Control Processes*. Princeton University Press, 1961.

[137] H. Zha, X. He, C. Ding, M. Gu, and H. Simon, "Spectral relaxation for k-means clustering," in *NIPS*, 2001, pp. 1057–1064.

[138] W. Labeeuw and G. Deconinck, "Non-intrusive detection of high power appliances in metered data and privacy issues," in *6th International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL 11)*, Copenhagen, Denmark, 2011.

[139] R. Presutto, M. Scialdoni, L. Cutaia, F. Lombardi, W. Mebane, R. Esposito, and S. Faberi, "Preparatory studies for eco-design requirements of EuPs, LOT 14: Domestic washing machines and dishwashers task 6 - 7," 2007, (Tender TREN/D1/40-2005).

[140] R. Kemna, W. van Elburg, Li, and R. van Holsteijn, "Methodology study eco-design of EuP, MEEUP product cases report," 2005, vHK for European Commission. [Online]. Available: http://ec.europa.eu/enterprise/policies/sustainable-business/ecodesign/methodology/files/finalreport2_en.pdf

[141] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *Security Privacy, IEEE*, vol. 7, no. 3, pp. 75 – 77, 2009.

[142] W. Labeeuw and G. Deconinck, "Customer sampling in a smart grid pilot," in *Power and Energy Society General Meeting, 2012 IEEE*, San Diego, USA, 2012.

[143] I. Lustig and J.-F. Puget, "Program does not equal program: Constraint programming and its relationship to mathematical programming," *Interfaces*, vol. 31, no. 6, pp. 29 – 53, 2001.

[144] K. Apt and M. Wallace, *Constraint Logic Programming using ECL$^i$PS$^e$*. Cambridge University Press, 2007.

[145] J. Löfberg, "Yalmip: A toolbox for modeling and optimization in MATLAB," in *Proceedings of the Computer-Aided Control System Design Conference, IEEE*, Taipei, Taiwan, 2004.

[146] K. Loquin and O. Strauss, "Fuzzy histograms and density estimation," in *Soft Methods for Integrated Uncertainty Modelling*, ser. Advances in Soft Computing. Springer Berlin / Heidelberg, 2006, vol. 37, pp. 45 – 52.

[147] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, 1st ed. Chapman and Hall / CRC, 1998.

[148] D. Woitrin and F. Possemiers, "Studie (F)111013-CDC-1113 over 'de geïnstalleerde capaciteit voor de productie van elektriciteit in belgië in 2010 en de evolutie ervan'," Commissie voor de Regulering van de Elektriciteit en het Gas (CREG), Tech. Rep., September 2011.

[149] Federal Public Service Economy, SMEs, Self-Employed and Energy, "De energiemarkt in 2009," 2011. [Online]. Available: http://statbel.fgov.be/nl/binaries/1585-11-01%20De%20energiemarkt%20in%202009_tcm325-140066.pdf

[150] D. Woitrin and F. Possemiers, "Beslissing (B)101223-CDC-1027 over 'de vraag tot goedkeuring van de evaluatiemethode voor en de bepaling van het primair, secundair en tertiair reservevermogen voor 2011'," Commissie voor de Regulering van de Elektriciteit en het Gas (CREG), Tech. Rep., December 2010.

[151] W. Labeeuw and G. Deconinck, "An architecture for collaborating data researchers in a smart grid pilot," in *Innovative Smart Grid Technologies (ISGT Europe), 2012 3rd IEEE PES International Conference and Exhibition on*, Berlin, 2012.

[152] M. Schumacher, "Firewall patterns," in *The 8th European Conference on Pattern Languages of Programs (EuroPLoP 2003)*, Irsee, Germany, June 2003.

[153] T. Ylonen and C. Lonvick, "The secure shell (SSH) transport layer protocol," *Request for Comments 4254*, 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4253.txt

[154] S. Chaudhuri and U. Dayal, "An overview of data warehousing and olap technology," *SIGMOD Rec.*, vol. 26, no. 1, pp. 65 – 74, 1997.

[155] H. Hasan and P. Hyland, "Using OLAP and multidimensional data for decision making," *IT Professional*, vol. 3, no. 5, pp. 44 – 50, Sep/Oct 2001.

[156] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: a warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1626 – 1629, Aug. 2009.

[157] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, "Pig latin: a not-so-foreign language for data processing," in *Proceedings of the*

*2008 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '08. New York, NY, USA: ACM, 2008, pp. 1099 – 1110.

[158] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin, "Hadoopdb: an architectural hybrid of mapreduce and dbms technologies for analytical workloads," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 922 – 933, Aug. 2009.

# Short CV

Wouter Labeeuw
Born on December 5th, 1984 in Kortrijk, Belgium

**1996 - 1998**
Secondary school
Heilig-Hartcollege Waregem
Moderne wetenschappen

**1998 - 2002**
Secondary school
Vrij Technisch Instituut (VTI) Waregem
Industriële wetenschappen

**2002 - 2005**
Katho - Vrij Hoger Technisch Instituut (VHTI) Kortrijk
Bachelor Elektronica - ICT

Bachelor's thesis: Processor gestuurd meetsysteem voor detectie interface

**2005 - 2007**
Howest - Provinciale Industriële Hogeschool (PIH) Kortrijk
Industrieel ingenieur elektronica, optie informatie- en communicatietechnologie

Master's thesis: Advanced graphics for Avionics Displays

**2007 - 2009**
Katholieke Universiteit Leuven (KU Leuven) Leuven
Burgerlijk ingenieur computerwetenschappen, optie artificiële intelligentie

Master's thesis: Samenwerkende agenten slaan alarm

**2009 - present**
Research assistant at KU Leuven, Department Electrical Engineering, ESAT-ELECTA

# List of publications

All publications are available at: http://www.esat.kuleuven.be/electa

### Reviewed Journals

Labeeuw, W.; Deconinck, G., "Residential Electrical Load Model Based on Mixture Model Clustering and Markov Models," Industrial Informatics, IEEE Transactions on, vol.9, no.3, pp.1561-1569, Aug. 2013

### International Conferences

Vencken Y., Labeeuw W., Deconinck G.: "Dealing with an Overdose of Photovoltaics at Distribution Level," PowerTech, Grenoble, France, June 16-20, 2013; 5 pages.

Labeeuw W., Deconinck G.: "An Architecture for Collaborating Data Researchers in a Smart Grid Pilot," IEEE PES ISGT, Berlin, October 14-17, 2012; 8 pages.

Labeeuw W., Deconinck G.: "Customer Sampling in a Smart Grid Pilot," IEEE PES GM'12, San Diego, USA, July 22-26, 2012; 7 pages.

Labeeuw W., Deconinck G.: "Non-intrusive detection of high power appliances in metered data and privacy issues," EEDAL 2011 - The 6th International Conference on Energy Efficiency in Domestic Appliances and Lighting, Copenhagen, Denmark, May 24-26, 2011; 7 pages.

Deconinck G., Labeeuw W., Vandael S., Beitollahi H., De Craemer K., Duan R., Qiu Z., Chittur Ramaswamy P., Vande Meerssche B., Vervenne I., Belmans R.: "Communication Overlays and Agents for Dependable Smart Power Grids," Proc. 5th CRIS Int. Conf. on Critical Infrastructures (CRIS-2010), Beijing, P.R.China, September 19-21, 2010; 7 pages.

Labeeuw W., Vandael S., Geth F, Deconinck G, "A MAS-based Smart Grid Simulator with PowerFlow-Analysis Integration," 5th IEEE Young Researchers Symposium (YRS2010), 2010; 4 pages.

Labeeuw W., Driessens K., Weyns D., , Deconinck G.: "Prediction of Congested Traffic on the Critical Density Point Using Machine Learning and Decentralised Collaborating Cameras," EPIA edition 2009, Aveiro, Portugal, October 12-15, 2009; pp. 15-26.

Bruynooghe M., De Cat B., Drijkoningen J., Fierens D., Goos J., Gutmann B., Kimmig A., Labeeuw W., Langenaken S., Landwehr N., Meert W., Nuyts E., Pellegrims R., Rymenants R., Segers S., Thon I., Van Eyck J., Van den Broeck G., Vangansewinkel T., Van Hove L., Vennekens J., Weytjens T., De Raedt L.: "An Exercise with Statistical Relational Learning Systems." In: Domingos, P., Kersting, K. (eds.) International Workshop on Statistical Relational Learning (SRL 2009), Leuven, Belgium, July 2-4, 2009; 3 pages.

FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL ENGINEERING
ELECTA
Kasteelpark Arenberg 10 box 2445
B-3001 Heverlee
wouter.labeeuw@esat.kuleuven.be
http://esat.kuleuven.be/electa